

Article

A Structural Labeling-File Audit and Secondary Model-Output Evaluation of Korean Specialized and Essential Medical Knowledge Datasets for Medical AI Research

Mi-ae Yang¹ and Kang-Su Ha^{2,3,*} 

¹ Department of Science and Technology Convergence, Graduate School, Chosun University, Gwangju 61452, Republic of Korea; bonbonheur@hanmail.net

² Department of Psychiatry, Rainbow Hospital, Gwangju 61924, Republic of Korea

³ College of Medicine, Chosun University, Gwangju 61452, Republic of Korea

* Correspondence: ksksha@naver.com

Abstract

Background: Korean-language medical question-answering datasets are increasingly used for large language model (LLM) development, but structural completeness alone does not establish model performance or clinical validity. We examined two national Korean medical knowledge datasets by combining a public labeling-file audit with a secondary evaluation of item-level outputs distributed with the corresponding official LLM packages. **Methods:** We reviewed the official documentation and complete public download inventories, parsed 34 training/validation labeling archives containing 31,036 records, and audited required fields, duplication, text length, question type, and specialty distribution. We also audited the official Qwen2.5-14B LoRA model packages and independently recalculated performance from their distributed item-level KorMedMCQA output files (2494 identical items per model). Accuracy was reported with Wilson 95% confidence intervals; model outputs were compared using an exact McNemar test and a paired bootstrap confidence interval. Because no compatible local GPU was available, model inference was not independently rerun. **Results:** The documentation described 34,487 labeled QA pairs, of which 31,036 (90.0%) were present in the publicly accessible training/validation labeling files; no public test-labeling archive was listed. Required fields were complete, no duplicated qa_id values were found, and one duplicated question-answer pair occurred in the Essential dataset. Multiple-choice items comprised 78.7% of all documented QA pairs, and the top three domains comprised 56.8%. In the distributed KorMedMCQA outputs, the Essential-care model answered 1603/2494 items correctly (64.27%; Wilson 95% CI 62.37–66.13%), while the Specialized-medicine model answered 1596/2494 correctly (63.99%; 95% CI 62.09–65.85%). The paired difference was 0.28 percentage points (bootstrap 95% CI −0.44 to 1.00), with no significant difference by exact McNemar test (46 vs. 39 discordant correct items; $p = 0.515$). **Conclusions:** The public training/validation labeling files showed favorable basic structural completeness, but the unavailable test labels, multiple-choice predominance, domain imbalance, limited record-level provenance, and incomplete model-package traceability constrain claims of clinical readiness. The distributed model outputs demonstrated moderate examination-style benchmark performance without a significant difference between the two models. These findings support use as research infrastructure, not evidence of clinical validity, and indicate the need for independent inference reproduction, clinician-led open-ended and safety evaluation, temporal updating, specialty-stratified reporting, and human oversight before clinical use.



Academic Editor: Susan McRoy

Received: 12 June 2026

Revised: 17 July 2026

Accepted: 21 July 2026

Published: 22 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

Keywords: structural dataset audit; secondary model-output evaluation; Korean medical knowledge; KorMedMCQA; medical question answering; large language models; model traceability; healthcare governance; patient safety

1. Introduction

Artificial intelligence (AI) and large language models (LLMs) are increasingly used in research on medical knowledge retrieval, question answering, clinical documentation support, medical education, and clinical decision support. However, the performance and safety of medical AI systems are not determined solely by model architecture. Dataset provenance, annotation design, specialty distribution, question format, currentness, bias, quality-control procedures, and reusability directly affect the reliability and clinical interpretability of downstream AI applications.

International benchmarks such as PubMedQA, MedQA, and MultiMedQA have enabled standardized comparisons of medical LLM performance [1–4]. Nevertheless, many such resources are built around English-language data, examination-style questions, research-abstract reasoning, or multiple-choice medical test formats. These features are useful for model benchmarking but may not adequately capture the open-ended communication, contextual reasoning, guideline currentness, and safety-critical decision-making required in real clinical settings.

Korean medical AI research requires resources that reflect Korean-language clinical and educational contexts. Korean medical documents frequently contain mixed Korean and English terminology, abbreviations, domestic specialty names, local documentation conventions, and expressions shaped by Korea's healthcare system. Directly applying English-language benchmarks to Korean medical AI development may therefore overlook important linguistic, cultural, and system-level characteristics.

AI Hub provides two public Korean medical knowledge resources: the Specialized Medical Knowledge Dataset and the Essential Medical Knowledge Dataset [5–8]. The former covers specialized domains such as surgery, psychiatry, and neurology/neurosurgery, whereas the latter focuses on Essential-care domains such as internal medicine, pediatrics/adolescent medicine, obstetrics/gynecology, and emergency medicine. Both datasets include JSON-based question-answering records and can serve as foundational resources for Korean medical natural language processing and LLM-based medical question-answering research.

Nevertheless, the existence of a large structured dataset does not establish clinical validity, safety, or effectiveness. LLM outputs may contain hallucinated information, outdated recommendations, incomplete safety warnings, overdiagnosis, under-triage, or domain-specific performance gaps. International guidance on AI for health also emphasizes transparency, human oversight, accountability, safety, and evaluation under realistic implementation conditions [9–14]. Therefore, medical knowledge datasets should be examined not only for scale but also for structural integrity, task readiness, error risks, and governance requirements before being used as infrastructure for clinical AI.

Recent studies and reviews have further described the translational challenges of medical AI, expanding applications of large language models in medical education and clinical support, automated clinical message classification, ethical responsibility, and primary-care integration [15–21].

The objective of this study was to conduct a structural labeling-file audit, a secondary audit of distributed model outputs, and a governance-focused analysis of the Korean Specialized Medical Knowledge Dataset and Essential Medical Knowledge Dataset. Specifically,

we aimed to (1) describe dataset architecture and public file coverage, (2) quantify question-format and specialty imbalance, (3) audit file-level integrity and documentation-to-file consistency, (4) evaluate the item-level KorMedMCQA outputs distributed with the official Qwen2.5-14B LoRA model packages, and (5) identify the evidentiary and governance requirements that remain before clinical use.

2. Materials and Methods

2.1. Study Design

This study combined a documentation-based structural analysis, a direct labeling-file audit, a model-package traceability audit, and a secondary statistical evaluation of distributed item-level benchmark outputs. It did not involve patient recruitment, clinical intervention, or assessment of patient outcomes. The empirical model component was limited to independent recalculation and paired analysis of output files distributed in the official Docker packages; the authors did not retrain the models or independently rerun inference.

2.2. Data Sources and Public File Scope

The analysis focused on the AI Hub Specialized Medical Knowledge Dataset and Essential Medical Knowledge Dataset, their description manuals and usage guidelines, the complete public download inventories available to the authors, labeling archives, model source code, adapter configurations, checkpoint state files, Docker images, and distributed model-output files. The public inventories contained 13 training and 13 validation labeling ZIP files for the Specialized dataset, and four training and four validation labeling ZIP files for the Essential dataset. No separate Test directory or test-labeling archive was listed. Files prefixed with TS were located under Training/Source data and were therefore treated as training-source archives rather than test sets.

2.3. Extracted Variables

From the dataset documentation, public inventories, and model packages, we extracted dataset scale, labeled QA counts, specialty domains, question types, JSON fields, documented split policy, file names and split locations, model base architecture, LoRA configuration, training-CSV size, checkpoint references, benchmark-output schema, and documentation/model-package inconsistencies. The distributed KorMedMCQA result files contained the question, the ground-truth option, the generated text, the extracted prediction, and the binary accuracy indicator for each item.

2.4. Documentation-Based and Labeling-File Data Audit

A documentation-based quantitative audit used the reported domain- and q_type-level counts. Domain concentration was summarized using the Herfindahl-Hirschman Index, $HHI = \sum_{i=1}^K p_i^2$, where p_i is the proportion of items in domain i ; the effective number of domains was $N_{\text{eff}} = 1/HHI$ [22]. HHI was used as a transparent concentration statistic rather than as a clinical quality score. The type-token ratio was retained only as a simple lexical-diversity proxy [23].

A file-level audit was performed on the two complete public labeling packages available to the authors, one per dataset. Training and validation labeling ZIP files were recursively parsed using Python and the standard JSON/ZIP processing workflow. For each JSON record, qa_id, domain, q_type, question, and answer were extracted; archive name and split were retained for traceability. A standalone standard-library parser implementing this recursive workflow is provided in Supplementary File S3. The parser accepts extracted directories and ZIP/TAR/TAR.GZ packages, recursively traverses nested labeling archives,

treats TS-prefixed files as training-source rather than test files, validates the five required QA fields, preserves source and split traceability, and reports malformed or nonconforming JSON objects rather than silently excluding them. Required-field missingness, exact duplicated identifiers, exact duplicated questions and question-answer pairs, character lengths, whitespace-delimited length proxies, short-answer frequency, and extreme long-answer thresholds were calculated. Parsed totals were reconciled with the public file inventory and documentation. Character counts were treated as the primary length measure. No Korean morphological analyzer was applied because the records contain mixed Korean-English medical terminology, and the audit prioritized a package-independent, exactly reproducible file-level proxy; whitespace-token length and type-token ratio were therefore not interpreted as Korean morphological or semantic quality measures.

2.5. Secondary Audit of Distributed Model Outputs

The official model packages identified Qwen/Qwen2.5-14B as the base model and LoRA adapters with rank 8, alpha 32, dropout 0.1, and query/key/value/output attention projections. The model-training CSVs contained 18,627 Essential-care items and 14,753 Specialized-medicine items; because these counts did not map directly to the publicly accessible training/validation labeling totals, the difference was recorded as a provenance limitation. The Docker packages contained precomputed item-level results for the same 2494 KorMedMCQA questions. We verified that question order and ground-truth labels were identical across the two output files and independently recalculated all performance statistics. The supplied evaluation scripts referenced `qwen_14b_essential_lora_cot_v2/checkpoint-32000` for the Essential-care output and `qwen_14b_professional_lora_cot_v2/checkpoint-28000` for the Specialized-medicine output. Checkpoint attribution was based on the scripts because checkpoint metadata were not embedded in the CSV files. This procedure is a secondary output audit, not an independent reproduction of inference.

2.6. Statistical Analysis

For each model, accuracy and a Wilson 95% confidence interval were calculated [24]. Because both files contained predictions for the same items, paired performance was compared using a two-sided exact McNemar test [25]. The absolute accuracy difference (Essential-care minus Specialized-medicine) was accompanied by a paired 95% bootstrap confidence interval based on 200,000 multinomial resamples with a fixed seed. We also calculated exact prediction agreement, enumerated discordant items, generated option-level confusion matrices, and repeated the comparison after retaining only the first occurrence of each exact duplicated benchmark question. Statistical reanalysis was performed in Python 3.13.5 with NumPy 2.3.5 and SciPy 1.17.0; the aggregate comparison workbook, reproducible paired-output audit script, and standalone labeling-file audit parser are supplied as Supplementary Files S1–S3. Two-sided $p < 0.05$ was considered statistically significant.

2.7. Task-Suitability and Governance Analysis

Structural suitability was evaluated separately for multiple-choice benchmarking, short-answer generation, descriptive medical response generation, retrieval-augmented generation, and specialty-stratified research. Potential risks were organized around provenance/currentness, specialty and question-format imbalance, unsupported generation, high-risk failures, human oversight, monitoring, and external validation. Governance recommendations were aligned with WHO transparency, human-oversight, and accountability principles, clinical AI reporting guidance, and Good Machine Learning Practice lifecycle principles [9–14,26]. No claim of ISO certification, medical-device conformity, or clinical deployment readiness was made.

2.8. Ethical Considerations

This study used publicly accessible, de-identified secondary datasets, documentation, legally obtained labeling files, model source files, and model-output files. It involved no new patient recruitment, clinical intervention, personally identifiable information, or patient-directed decision-making. Individual training QA records were not redistributed in the manuscript. The KorMedMCQA output audit used examination-style benchmark questions and model predictions, not patient data. Therefore, the study was considered not to constitute human-subjects or human-materials research.

3. Results

3.1. Overall Dataset Architecture and Scale

Both datasets are text-based question-answering resources for Korean medical AI research. The Specialized dataset includes 13 specialized domains, whereas the Essential dataset includes four essential-care domains. Together, the two datasets contain 222,073,009 source-data tokens and 34,487 labeled QA pairs according to the dataset documentation. Figure 1 summarizes the analytic framework used in this study.

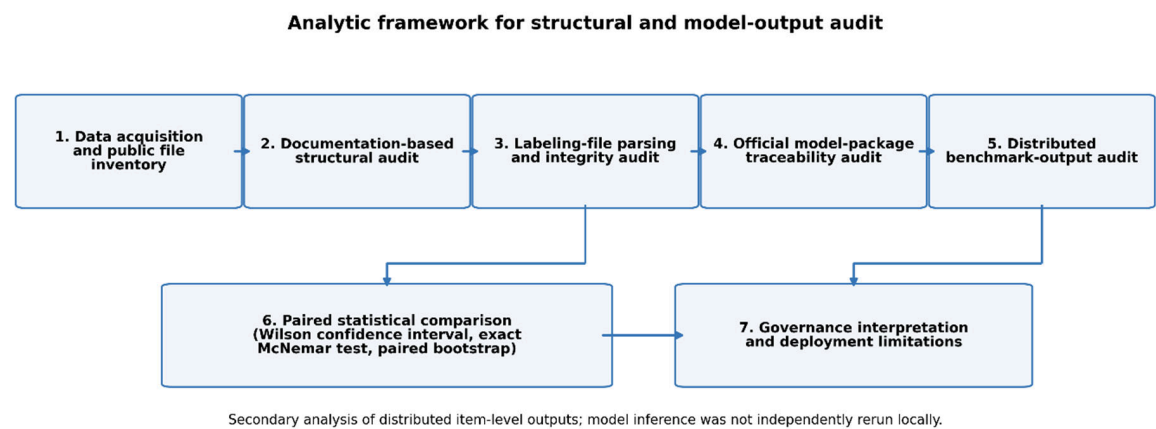


Figure 1. Analytic framework for the structural labeling-file audit, official model-package traceability audit, secondary distributed-output evaluation, and governance interpretation.

Figure 2 summarizes the paired secondary analysis of 4988 distributed item-level outputs (2494 per model), including model accuracy with Wilson 95% confidence intervals and the concordant/discordant response pattern. Detailed statistics are reported in Section 3.5.

The detailed characteristics of the two datasets are summarized in Table 1.

Table 1. Overview and technical characteristics of the two Korean medical knowledge datasets.

Item	Specialized Medical Knowledge Dataset	Essential Medical Knowledge Dataset	Integrated Interpretation
English name	Specialized Medical Knowledge	Essential Medical Knowledge	Korean medical knowledge QA resources
Data field	Data	Data	Medical AI research infrastructure
Data type	Healthcare	Healthcare	Natural-language processing applicable
File format	Text	Text	Structured processing applicable
Labeling type	JSON	JSON	QA tasks applicable
Source-data scale	Natural-language QA	Natural-language QA	222,073,009 tokens
Labeled QA scale	120,673,006 tokens	101,400,003 tokens	34,487 pairs
Medical domains	15,286 pairs	19,201 pairs	17 nominal domains
	13 domains	4 domains	
Question types	Multiple-choice, short-answer, descriptive	Multiple-choice, short-answer, descriptive	Question-type-specific evaluation possible
Split policy	8:1:1	8:1:1	Training/validation/test reproducibility

Note: Prepared by the authors based on AI Hub dataset documentation and usage guidelines [5–8].

Paired secondary audit of distributed KorMedMCQA outputs (n = 2494 items per model)

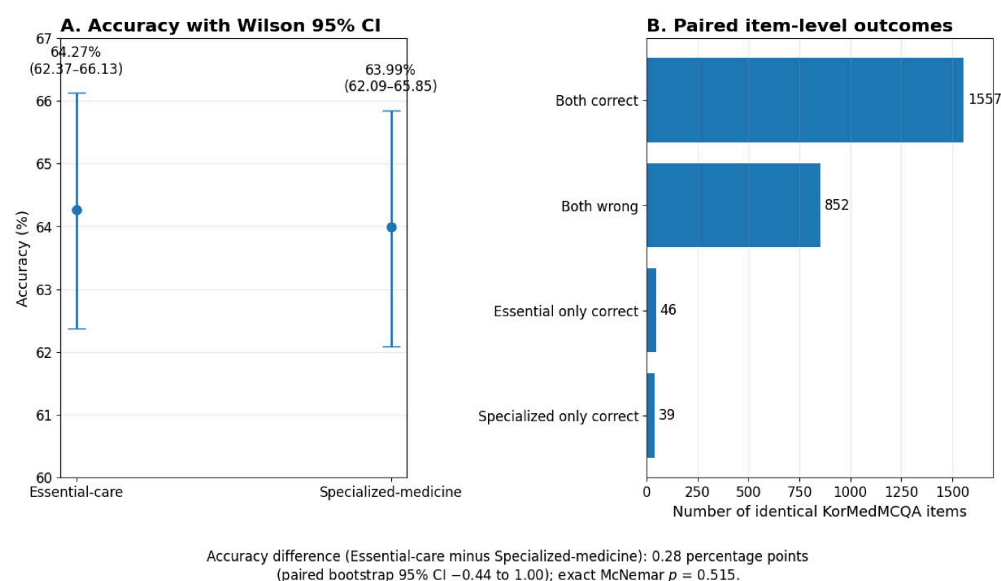


Figure 2. Paired secondary analysis of 4988 distributed KorMedMCQA item-level outputs (2494 per model). Panel (A) shows accuracy with Wilson 95% confidence intervals. Panel (B) shows paired concordant and discordant outcomes across the 2494 identical questions; the exact McNemar test yielded $p = 0.515$.

3.2. JSON Annotation Structure

Both datasets use a common JSON-based QA structure. The key fields are `qa_id`, `domain`, `q_type`, `question`, and `answer`. This structure supports item tracking, domain-stratified analysis, question-type-specific evaluation, and linkage of input questions to reference answers (Table 2).

Table 2. Key JSON fields in the labeled QA records.

Field	Data Type	Meaning	Technical Utility
qa_id	integer or identifier	Unique identifier for each QA record	Duplication check, item tracking, split verification
domain	integer code	Medical specialty or domain	Domain-specific analysis and benchmarking
q_type	integer code	Question type: 1 = multiple-choice, 2 = short-answer, 3 = descriptive	Question-format-specific evaluation
question	text	Input medical question	Model input, retrieval query, prompt
answer	text	Reference answer	Model-output comparison, reference response

3.3. Domain and Question-Type Distribution

The largest domain was Essential-Internal medicine (12,781 QA pairs; 37.1%), followed by Specialized-Surgery (3731; 10.8%) and Essential-Pediatrics/adolescent medicine (3087; 9.0%). The top three domains accounted for 56.8% of all labeled QA pairs, and the HHI for domain distribution was 0.178, corresponding to an effective number of domains of 5.62. This indicates substantial domain concentration despite the nominal presence of 17 domains (Table 3 and Figure 3).

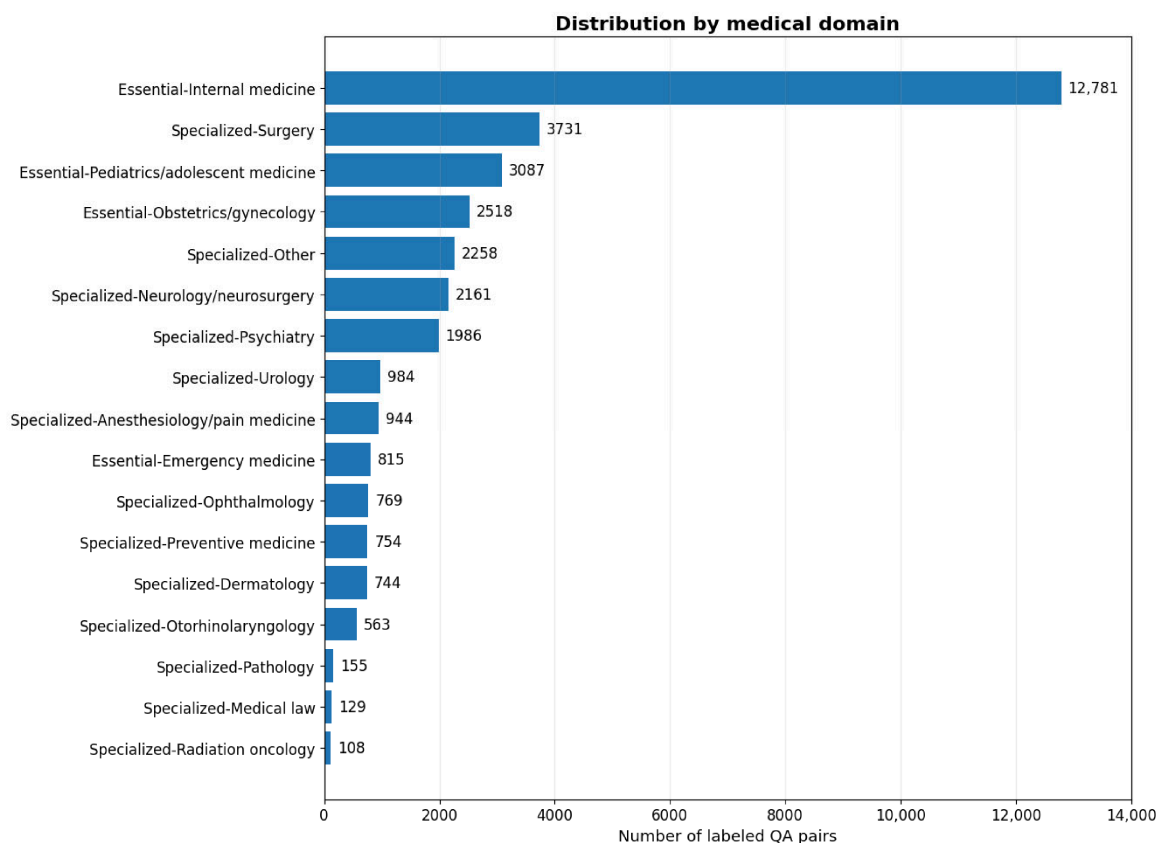


Figure 3. Distribution of labeled QA pairs by medical domain.

Table 3. Medical-domain and question-type distribution.

Category	Item	Number of Labeled QA Pairs	Pooled Percentage
Domain	Essential-Internal medicine	12,781	37.1%
Domain	Specialized-Surgery	3,731	10.8%
Domain	Essential-Pediatrics/adolescent medicine	3,087	9.0%
Domain	Essential-Obstetrics/gynecology	2,518	7.3%
Domain	Specialized-Other	2,258	6.5%
Domain	Specialized-Neurology/neurosurgery	2,161	6.3%
Domain	Specialized-Psychiatry	1,986	5.8%
Domain	Specialized-Urology	984	2.9%
Domain	Specialized-Anesthesiology/pain medicine	944	2.7%
Domain	Essential-Emergency medicine	815	2.4%
Domain	Specialized-Ophthalmology	769	2.2%
Domain	Specialized-Preventive medicine	754	2.2%
Domain	Specialized-Dermatology	744	2.2%
Domain	Specialized-Otorhinolaryngology	563	1.6%
Domain	Specialized-Pathology	155	0.4%

Table 3. Cont.

Category	Item	Number of Labeled QA Pairs	Pooled Percentage
Domain	Specialized-Medical law	129	0.4%
Domain	Specialized-Radiation oncology	108	0.3%
q_type	Multiple-choice	27,156	78.7%
q_type	Short-answer	3610	10.5%
q_type	Descriptive	3721	10.8%

Multiple-choice questions accounted for 27,156 QA pairs (78.7%), whereas short-answer and descriptive questions together accounted for 7331 pairs (21.3%). The ratio of multiple-choice to open-ended items was approximately 3.7:1. This structure is advantageous for automated scoring and benchmark construction but may not fully represent the open-ended complexity of real clinical question answering.

3.4. Labeling-File Data Audit Results

The file-level audit parsed 31,036 QA records from all 34 publicly listed training/validation labeling ZIP files: 13,756 records from 26 Specialized files and 17,280 records from eight Essential files. These totals represented 90.0% of the 34,487 documented QA pairs. The complete public inventories contained no Test directory or test-labeling archive; the numerical remainders (1530 Specialized and 1921 Essential records) were consistent with the documented 10% test allocation but could not be directly audited.

No missing question, answer, or qa_id fields and no duplicated qa_id values were observed. One duplicated question and one identical question-answer pair were detected in the Essential dataset. The pooled mean question length was 198.87 ± 89.55 characters, and the pooled mean answer length was 46.96 ± 83.53 characters. Whitespace-based measures are reported only as reproducible lexical proxies: pooled question length 45.37 ± 19.23 units, pooled answer length 11.14 ± 18.96 units, and 15,716 answers (50.6%) shorter than five units. The detailed results of the labeling-file data audit are summarized in Table 4.

Table 4. Results of the labeling-file data audit.

Audit Item	Specialized Dataset	Essential Dataset	Integrated Interpretation
Audit file scope	26 training/validation labeling ZIP files	8 training/validation labeling ZIP files	34 ZIP.part0 labeling files in total
Parsed QA records	13,756	17,280	31,036; 90.0% of the documented total of 34,487 pairs
Test set included	No	No	The available TAR files did not contain test-set labeling files
Missing question	0 (0.0%)	0 (0.0%)	0 (0.0%)
Missing answer	0 (0.0%)	0 (0.0%)	0 (0.0%)
Missing qa_id	0 (0.0%)	0 (0.0%)	0 (0.0%)
Duplicated qa_id	0	0	0
Duplicated question	0	1	1 in total
Identical question-answer pair	0	1	One record in the internal medicine domain of the Essential dataset
Question character length	192.84 ± 86.46 ; median 180	203.68 ± 91.65 ; median 190	198.87 ± 89.55 ; median 185
Answer character length	51.85 ± 89.30 ; median 18	43.08 ± 78.41 ; median 17	46.96 ± 83.53 ; median 17
Question whitespace-token length	44.62 ± 19.18	45.98 ± 19.24	45.37 ± 19.23
Answer whitespace-token length	12.30 ± 20.31	10.23 ± 17.76	11.14 ± 18.96
Answers shorter than five tokens	6627 (48.2%)	9089 (52.6%)	15,716 (50.6%)
Top 1% long-answer threshold	≥ 89 tokens; 139 records	≥ 86 tokens; 175 records	≥ 87 tokens; 316 records
Whitespace-based type-token ratio	0.09907	0.08203	0.07244

Note: The table summarizes direct parsing of the available AI Hub training/validation labeling ZIP.part0 files. Because the available TAR files did not contain test-set labeling files, the audit should be interpreted as covering 31,036 of 34,487 labeled QA pairs (90.0%).

3.5. Secondary Audit of Distributed KorMedMCQA Outputs

Both distributed output files contained all 2494 KorMedMCQA items, and no predictions were missing or unparsable. The Essential-care model had 1603 correct responses (64.27%; Wilson 95% CI 62.37–66.13%), whereas the Specialized-medicine model had 1596 correct responses (63.99%; 95% CI 62.09–65.85%). Both models were correct on 1557 items and wrong on 852 items; the Essential-care model alone was correct on 46 items and the Specialized-medicine model alone on 39 items. The exact McNemar test showed no significant difference ($p = 0.515$). The absolute accuracy difference was 0.28 percentage points, with a paired bootstrap 95% CI of -0.44 to 1.00 percentage points. The models selected the same response option on 2364 items (94.79%) and different options on 130 items. After exact duplicate questions were collapsed to 2489 unique questions, results were materially unchanged (64.24% vs. 64.00%; exact McNemar $p = 0.586$). Table 5 summarizes the paired output audit.

Table 5. Secondary audit and paired comparison of distributed KorMedMCQA model outputs.

Metric	Essential-Care Model	Specialized-Medicine Model	Paired Interpretation
Evaluated items	2494	2494	Identical questions and ground truth
Correct responses	1603	1596	7 more correct for Essential-care
Accuracy	64.27%	63.99%	Difference 0.28 percentage points
Wilson 95% CI	62.37–66.13%	62.09–65.85%	Substantially overlapping intervals
Missing/unparsable predictions	0	0	Complete extracted predictions
Both models correct	1557	Same paired count	Paired concordant-correct cell
Both models wrong	852	Same paired count	Paired concordant-wrong cell
Only this model correct	46	39	Discordant cells
Exact McNemar test	-	-	$p = 0.515$ (two-sided)
Paired bootstrap 95% CI	-	-	Accuracy difference -0.44 to 1.00 percentage points
Same predicted option	2364 (94.79%)	Same paired count	High option-level agreement
Different predicted option	130 (5.21%)	Same paired count	Item-level discordance retained in Supplementary File S1

3.6. Structural Task Suitability and Research Use

The common JSON schema, explicit domain labels, and q_type labels make the datasets structurally usable for supervised QA research, specialty-stratified benchmarking, and question-format-specific analyses. The distributed model outputs provide empirical evidence that the resources have been used in examination-style multiple-choice modeling. However, approximately 64% KorMedMCQA accuracy does not demonstrate safe open-ended generation, diagnostic accuracy, treatment appropriateness, or patient benefit. Suitability should therefore be stated at the level of research infrastructure and task format, not clinical deployment.

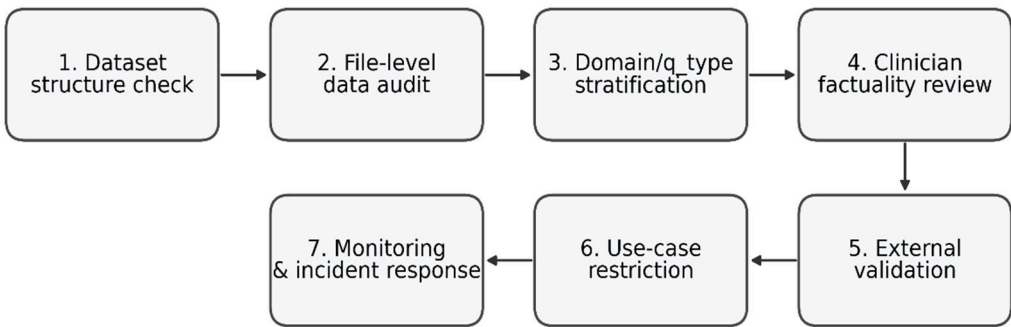
3.7. Error Risks and Governance Requirements

The structural and model-package audits identified risks related to unavailable public test labels, record-level source/currentness gaps, specialty imbalance, multiple-choice dominance, model-training CSV counts that were not directly mapped to public labeling files, incomplete model cards, and output files lacking embedded checkpoint and hardware metadata. Table 6 organizes these risks and controls using established trans-

parency, lifecycle monitoring, human oversight, and external validation principles [9–14,26]. Figure 4 illustrates a pre-implementation governance workflow.

Table 6. Error risks, evaluation criteria, and governance requirements for medical AI research and any future clinical evaluation.

Analytic Area	Potential Risk	Clinical Implication	Control Measure
Data provenance/currentness	Different source dates; delayed guideline updates	Outdated recommendations may be generated	Maintain source dates, guideline versions, and update schedules
Domain distribution	Concentration in selected domains	Domain-specific performance imbalance	Report performance by specialty and reinforce low-frequency domains
Question type	Multiple-choice predominance	Limited representation of real clinical dialogue	Evaluate short-answer and descriptive items separately using expert rubrics
LLM generation	Hallucination, inaccurate explanation, unsupported treatment suggestions	Patient-safety risk	Link outputs to source documents, factuality assessment, retrieval-augmented generation, citation display
High-risk situations	Failure to detect emergency signs or contraindications	Potential harm or delay in care	High-risk item checklist and clinician review
Governance	Unclear responsibility; automation bias	Legal and ethical issues in clinical deployment	Multidisciplinary committee, audit trail, and use-case restriction
Evaluation rubric	Unclear criteria for assessing responses	Difficult evaluation of descriptive answers	Score medical factuality, currentness, safety, hallucination risk, and applicability on a 1–5 scale
External validation	Reliance on internal data only	Limited generalizability	External Korean medical QA validation and inter-institutional evaluation



Multidisciplinary Review Committee: clinicians, AI engineers, data-governance experts, and ethics/legal experts

Figure 4. Safety and governance framework before clinical AI deployment.

4. Discussion

4.1. Principal Findings and Model-Output Interpretation

This study combined a complete audit of the publicly accessible training/validation labeling files with a secondary audit of the item-level KorMedMCQA outputs distributed in the two official model packages. The labeling files showed high basic structural completeness, while the output files showed moderate examination-style accuracy: 64.27% for the Essential-care model and 63.99% for the Specialized-medicine model.

The paired analysis is more informative than comparing two rounded percentages. The models agreed on the selected option for 94.79% of items, and only 85 items were correct for one model but not the other. The exact McNemar result and paired bootstrap interval did not support a meaningful performance difference. The outputs therefore provide evidence of similar benchmark behavior, not evidence that either model is clinically safe or superior for patient care.

4.2. Question Format, Domain Distribution, and Benchmark Implications

KorMedMCQA and the source datasets are predominantly multiple-choice or examination-style resources. These formats support reproducible scoring but underrepresent history-taking, clarification, uncertainty communication, contraindication checking, guideline citation, and longitudinal reasoning. When model outputs are exposed to implementation settings, time pressure, incomplete context, and user overreliance can amplify the practical consequences of otherwise small benchmark errors. Accordingly, the observed benchmark accuracy should not be generalized to open-ended clinical conversations or clinical deployment exposure.

The datasets also contain marked specialty imbalance. Internal medicine accounts for 37.1% of all documented QA pairs, and the top three domains account for 56.8%. Such concentration can increase uncertainty in low-frequency specialties and make pooled accuracy insensitive to weak performance in radiation oncology, medical law, pathology, or other small domains.

Future evaluations should therefore use prespecified specialty strata, balanced sampling or weighting, per-domain uncertainty intervals, and separate reporting by question type.

Mixed Korean-English terminology, local specialty coding, and heterogeneous source dates create additional sources of language and temporal biases. Record-level source links and guideline versions were not available in the labeling schema, so individual-answer currentness could not be verified. These limitations are relevant to other local-language public medical datasets and reinforce the need to audit provenance and update processes rather than relying on dataset size alone.

4.3. Governance and Model-Package Traceability

A strength of the resources is the explicit JSON question-answer schema and the availability of model code, adapters, checkpoint state files, Docker images, and item-level output files. These materials enabled the present traceability audit and independent recalculation of benchmark statistics.

Important traceability gaps remain. Public training/validation labeling totals did not directly match model-training CSV counts, and the distributed result CSV files did not embed checkpoint, hardware, execution-time, or random-seed metadata. The generated model cards also left key training-data, intended-use, risk, and evaluation fields largely incomplete.

Although the supplied run scripts linked the outputs to checkpoint-32000 and checkpoint-28000, this linkage could not be verified from the result files alone. These findings support lifecycle documentation, immutable manifests, model and dataset version identifiers, cryptographic checksums, executable environment capture, and preservation of successful run logs.

Governance recommendations were therefore organized around WHO principles, clinical AI reporting guidance, and Good Machine Learning Practice [9–14,26]. They include explicit intended-use restrictions, record-level provenance, change control, specialty- and risk-stratified evaluation, independent reproduction, human review of open-ended outputs, post-deployment monitoring, and accountability for model updates.

4.4. Recommendations for Researchers and Dataset Curators

Researchers should describe the exact public file inventory, split availability, parsing environment, model checkpoint, prompt, decoding settings, and output-extraction rules. Benchmark reports should include uncertainty intervals, paired comparisons when models answer the same items, parsing-failure counts, duplicate-item sensitivity analyses, and domain/question-type disaggregation.

Open-ended outputs require clinician-developed rubrics for factuality, guideline consistency, omission, hallucination, contraindications, and high-risk escalation.

Dataset curators should publish a machine-readable manifest linking each public record to its split, source-document identifier, source date, guideline version, annotation history, and quality-control status. Public model packages should include a completed model card, exact training-record manifest, environment lock file, checkpoint selection rationale, successful evaluation log, and item-level outputs with embedded provenance.

Low-frequency specialties and open-ended formats should be expanded, and obsolete medical knowledge should be managed through scheduled review and versioned updates.

Any patient-facing or decision-support use should remain outside the evidence supported by this study. Before such use, independent model execution, external Korean clinical validation, specialist adjudication, prospective workflow evaluation, monitoring for distribution shift, and human-in-the-loop escalation are required.

4.5. Limitations

First, the labeling-file audit covered all publicly accessible training/validation files but not the unavailable test-labeling files; therefore, it cannot establish completeness or quality of the full documented dataset. Second, the study did not independently adjudicate the clinical correctness, guideline currentness, contraindications, or safety of the reference answers or generated explanations.

Third, the model component was a secondary audit of item-level outputs included in official Docker images; inference was not independently rerun, and checkpoint attribution depended on the supplied scripts rather than metadata embedded in the result files. Fourth, KorMedMCQA is an examination-style multiple-choice benchmark and does not measure open-ended clinical communication, retrieval fidelity, hallucination, or patient outcomes.

Fifth, the model-training CSV counts could not be mapped record-by-record to the public labeling archives, limiting leakage and provenance verification beyond exact question matching. Sixth, the benchmark output files did not contain specialty labels, preventing specialty-stratified performance analysis. Seventh, whitespace-token counts and type-token ratio are simple mixed-language proxies rather than Korean morphological measures. Eighth, five exact duplicate benchmark questions were identified; a deduplicated sensitivity analysis did not change the paired conclusion. Finally, the analysis evaluated only two released model variants and should not be generalized to other models or deployment settings.

5. Conclusions

The Korean Specialized Medical Knowledge Dataset and Essential Medical Knowledge Dataset provide substantial, structured resources for Korean medical AI research. All publicly listed training/validation labeling files were audited, and the required fields showed favorable basic completeness. A secondary audit of distributed KorMedMCQA outputs found moderate accuracy for both released Qwen2.5-14B LoRA models (64.27% and 63.99%) and no statistically significant paired difference.

These findings do not establish clinical AI readiness. Public test labels were unavailable, question formats and specialties were imbalanced, record-level source/currentness metadata were limited, model-training-set provenance was incomplete, and inference was not independently reproduced. The datasets and model packages are best interpreted as research infrastructure that requires stronger manifests, versioning, independent execution, open-ended clinician evaluation, safety testing, external validation, continuous updating, and human oversight before any clinical application.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/biomedinformatics6040049/s1>, Supplementary File S1: KorMedMCQA paired model-output comparison workbook, including summary statistics, option-level performance, confusion matrices, discordant items, and duplicate-question sensitivity analysis. Supplementary File S2: Reproducible paired-output audit script. Supplementary File S3: Standalone standard-library labeling-file audit parser for extracted directories and ZIP/TAR packages, including nested-archive traversal, required-field validation, source/split traceability, duplicate checks, text-length statistics, and explicit issue reporting.

Author Contributions: Conceptualization, M.-a.Y. and K.-S.H.; methodology, M.-a.Y.; data curation and labeling-file audit, M.-a.Y.; model-package traceability and statistical analysis, M.-a.Y.; clinical interpretation and governance review, K.-S.H.; writing—original draft preparation, M.-a.Y.; writing—review and editing, M.-a.Y. and K.-S.H.; supervision and correspondence, K.-S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was a technical data audit of publicly available, de-identified secondary datasets, dataset documentation, usage guidelines, and legally obtained labeling files. The study did not involve new patient recruitment, clinical intervention, collection of personally identifiable information, individual-level patient data analysis, or patient-directed clinical decision-making. Only aggregate-level results were reported. Therefore, this study was considered not to constitute human-subjects research or human-materials research, and a separate IRB approval number was not applicable.

Informed Consent Statement: Not applicable. This study did not involve human participants, patient recruitment, clinical intervention, collection of personally identifiable information, or individual-level patient data analysis.

Data Availability Statement: Restrictions apply to the underlying AI Hub datasets, which are available according to AI Hub access and usage policies. Original source documents and full labeling records are not redistributed. The aggregate structural audit, paired KorMedMCQA comparison workbook, paired-analysis script, and labeling-file audit parser are provided as Supplementary Materials. The item-level model-output files were obtained from the official AI Hub Docker distributions and are subject to the applicable package and dataset terms.

Acknowledgments: During preparation of this manuscript, generative AI tools were used for language refinement, structural organization, and formatting assistance. These tools were not used to make clinical judgments, fabricate data, or replace author-led interpretation of the labeling-file audit results. The authors reviewed and edited all AI-assisted outputs and take full responsibility for the integrity, accuracy, and final content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W.W.; Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 2567–2577. [\[CrossRef\]](#)
2. Jin, D.; Pan, E.; Oufattole, N.; Weng, W.H.; Fang, H.; Szolovits, P. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *arXiv* **2020**, arXiv:2009.13081. [\[CrossRef\]](#)
3. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large Language Models Encode Clinical Knowledge. *Nature* **2023**, *620*, 172–180. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Nori, H.; King, N.; McKinney, S.M.; Carignan, D.; Horvitz, E. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv* **2023**, arXiv:2303.13375. [\[CrossRef\]](#)
5. AI Hub. 2024 *Specialized Medical Knowledge Dataset Description Manual, Version 03*; National Information Society Agency: Daegu, Republic of Korea, 2025. Available online: <https://www.aihub.or.kr/aihubdata/data/view.do?aihubDataSe=data&dataSetSn=71874> (accessed on 17 April 2026).

6. AI Hub. 2024 *Essential Medical Knowledge Dataset Description Manual, Version 03*; National Information Society Agency: Daegu, Republic of Korea, 2025. Available online: <https://www.aihub.or.kr/aihubdata/data/view.do?aihubDataSe=data&dataSetSn=71875> (accessed on 17 April 2026).
7. AI Hub. 2024 *Specialized Medical Knowledge Usage Guidelines, Version 5.0*; National Information Society Agency: Daegu, Republic of Korea, 2025. Available online: <https://www.aihub.or.kr/aihubdata/data/view.do?aihubDataSe=data&dataSetSn=71874> (accessed on 17 April 2026).
8. AI Hub. 2024 *Essential Medical Knowledge Usage Guidelines, Version 5.0*; National Information Society Agency: Daegu, Republic of Korea, 2025. Available online: <https://www.aihub.or.kr/aihubdata/data/view.do?aihubDataSe=data&dataSetSn=71875> (accessed on 17 April 2026).
9. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*; World Health Organization: Geneva, Switzerland, 2021. Available online: <https://www.who.int/publications/i/item/9789240029200> (accessed on 17 April 2026).
10. World Health Organization. *Regulatory Considerations on Artificial Intelligence for Health*; World Health Organization: Geneva, Switzerland, 2023. Available online: <https://www.who.int/publications/i/item/9789240078871> (accessed on 17 April 2026).
11. Vasey, B.; Nagendran, M.; Campbell, B.; Clifton, D.A.; Collins, G.S.; Denaxas, S.; Denniston, A.K.; Faes, L.; Geerts, B.; Ibrahim, M.; et al. Reporting Guideline for the Early-Stage Clinical Evaluation of Decision Support Systems Driven by Artificial Intelligence: DECIDE-AI. *Nat. Med.* **2022**, *28*, 924–933. [[CrossRef](#)] [[PubMed](#)]
12. Collins, G.S.; Moons, K.G.M.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; van Smeden, M.; et al. TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *BMJ* **2024**, *385*, e078378. [[CrossRef](#)] [[PubMed](#)]
13. Liu, X.; Cruz Rivera, S.; Moher, D.; Calvert, M.J.; Denniston, A.K. SPIRIT-AI and CONSORT-AI Working Group. Reporting Guidelines for Clinical Trial Reports for Interventions Involving Artificial Intelligence: The CONSORT-AI Extension. *Nat. Med.* **2020**, *26*, 1364–1374. [[CrossRef](#)] [[PubMed](#)]
14. Cruz Rivera, S.; Liu, X.; Chan, A.W.; Denniston, A.K.; Calvert, M.J. SPIRIT-AI and CONSORT-AI Working Group. Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence: The SPIRIT-AI Extension. *BMJ* **2020**, *370*, m3210. [[CrossRef](#)] [[PubMed](#)]
15. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; King, D. Key Challenges for Delivering Clinical Impact with Artificial Intelligence. *BMC Med.* **2019**, *17*, 195. [[CrossRef](#)] [[PubMed](#)]
16. Rajpurkar, P.; Chen, E.; Banerjee, O.; Topol, E.J. AI in Health and Medicine. *Nat. Med.* **2022**, *28*, 31–38. [[CrossRef](#)] [[PubMed](#)]
17. Vrdoljak, J.; Boban, Z.; Vilovic, M.; Kumric, M.; Bozic, J. A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration. *Healthcare* **2025**, *13*, 603. [[CrossRef](#)] [[PubMed](#)]
18. Lorenzoni, G.; Garbin, A.; Brigiari, G.; Papappicco, C.A.M.; Manfrin, V.; Gregori, D. Large Language Models in Action: Supporting Clinical Evaluation in an Infectious Disease Unit. *Healthcare* **2025**, *13*, 879. [[CrossRef](#)] [[PubMed](#)]
19. Shifa, A.; Ali, G.G.M.N.; Foulger, R. Leveraging Large Language Models for Automating Outpatients' Message Classifications of Electronic Medical Records. *Healthcare* **2025**, *13*, 3052. [[CrossRef](#)] [[PubMed](#)]
20. Martinho, D.; Sobreiro, P.; Domingues, A.; Martinho, F.; Nogueira, N. Ethical Responsibility in Medical AI: A Semi-Systematic Thematic Review and Multilevel Governance Model. *Healthcare* **2026**, *14*, 287. [[CrossRef](#)] [[PubMed](#)]
21. Yousefi, F.; Dehnavieh, R.; Laberge, M.; Haghdoust, A.; Sasseville, M.; Hekmat, S.; Ghaemi, M.; Nadali, M. Integrating Artificial Intelligence (AI) in Primary Health Care (PHC) Systems: A Framework-Guided Comparative Qualitative Study. *Healthcare* **2026**, *14*, 145. [[CrossRef](#)] [[PubMed](#)]
22. Jost, L. Entropy and Diversity. *Oikos* **2006**, *113*, 363–375. [[CrossRef](#)]
23. Baayen, R.H. *Word Frequency Distributions*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 2001; Volume 18, ISBN 978-0-7923-7017-8. [[CrossRef](#)]
24. Brown, L.D.; Cai, T.T.; DasGupta, A. Interval Estimation for a Binomial Proportion. *Stat. Sci.* **2001**, *16*, 101–133. [[CrossRef](#)]
25. McNemar, Q. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157. [[CrossRef](#)] [[PubMed](#)]
26. U.S. Food and Drug Administration; Health Canada; Medicines and Healthcare Products Regulatory Agency. Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021. Available online: <https://www.fda.gov/media/153486/download> (accessed on 17 April 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.