

Knowledge Distillation-Based Lightweight Generative Large Language Models for Standardized Quality Control of Chinese Radiology Reports

Duoning Jiang

Anhui University of Chinese Medicine

Xiao Han

Anhui Imaging Union Cloud Medical Technology Co., Ltd

Ke Xu

Anhui Imaging Union Cloud Medical Technology Co., Ltd

Chuanfu Li

l i c f @ a h t c m . e d u . c n

The First Affiliated Hospital of Anhui University of Chinese Medicine

Article

Keywords: Large language model, Local deployment, Chinese Radiology Reports, Standardized Quality Control, knowledge distillation

Posted Date: September 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10850457/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on October 16th, 2026. See the published version at <https://doi.org/10.1038/s41598-026-76600-y>.

Abstract

Radiological reports, which document radiologists' imaging observations and diagnostic conclusions from CT, X-ray and MRI scans, deliver core imaging evidence for clinical diagnosis and therapeutic planning. Inaccurate or disordered reports may lead to misdiagnosis and delayed intervention, posing risks to patient safety. Radiological reports are essential for clinical decision-making, yet Chinese radiology reports frequently exhibit logical inconsistencies, incomplete content, non-standardized descriptions, and formatting errors, while manual quality control remains time-consuming and inconsistent. Large teacher models can provide task-specific supervision but are difficult to deploy locally, whereas smaller models require adaptation to this clinical text task. We developed a teacher-supervised, locally deployable 4B language-model system for rule-based textual quality control of Chinese radiology reports. Among the three evaluated teacher APIs, Doubao-seed1.6-non-thinking achieved the highest macro-precision 0.94, under the study's precision-prioritized selection. Qwen3-4B-Instruct-2507 emerged as the optimal student model, achieving an average precision of 0.90, recall of 0.49, and F1 score of 0.62, with inference speed improved by over 40-fold. Real-world clinical validation demonstrated that the lightweight model attains a precision of 0.74 in authentic clinical scenarios while retaining high overall performance, satisfying the deployment demands of on-site clinical institutions.

INTRODUCTION

Radiological reports serve as critical textual foundations for clinical diagnosis and treatment decision-making, and the standardization of report writing directly impacts diagnostic accuracy, clinical efficiency, and patient safety¹⁻³. Chinese radiology reports frequently exhibit issues such as logical inconsistencies, incomplete content, non-standardized lesion descriptions, and textual and formatting errors in clinical practice. Traditional manual quality control (QC) methods^{4,5} are time-consuming, labor-intensive, and lack consistency, making large-scale standardized quality management difficult to achieve.

In recent years, generative LLMs have demonstrated significant potential in medical text processing and QC, effectively identifying various types of errors in reports⁶⁻¹⁰. However, full-parameter large models have limitations such as massive parameter sizes, high deployment costs, slow inference speeds, and data privacy risks, making them difficult to deploy locally in hospitals¹¹⁻¹³. Lightweight small models, while easy to deploy, lack sufficient medical domain knowledge and cannot meet clinical QC accuracy requirements^{14,15}.

Teacher-generated labels or responses can be used to fine-tune smaller models, potentially reducing inference requirements, and knowledge distillation can transfer professional capabilities from large models to lightweight small models, substantially reducing resource consumption while maintaining model performance, providing a feasible solution to the aforementioned problems. Woo et al. utilized synthetic knowledge distillation to build lightweight clinical text extractors from large foundation models. These compact models retain comparable analytical performance for radiology and clinical notes, delivering 4-fold faster inference speed and cutting cloud deployment costs by over 75%, thus bringing far lower computational overhead for routine clinical deployment¹⁶. However, existing studies mainly rely on retrospective data and preset errors, with limited clinical validation¹⁷. The use of knowledge-distilled lightweight LLMs for standardized QC of Chinese radiology reports also remains underexplored, lacking systematic model selection, distillation, and clinical evaluation.

For radiology-report QC, previous deployment strategies are divided into Cloud-only Deployment and Local-only Deployment. As shown in Fig. 1A, the Cloud-only Deployment employs proprietary cloud-hosted LLMs for report processing yet causes risks of off-site medical-data transmission and privacy disclosure. As shown in Fig. 1B, the Local-only Deployment runs lightweight models entirely on-premise, while limited local computing resources restrict model performance. Our scheme is displayed in Fig. 1C, the Hybrid Deployment combining cloud-side teacher-model training and hospital-local student-model inference, which addresses the shortcomings of the two conventional patterns. On this basis, we construct a teacher-supervised lightweight system following thirteen text-checking rules for Chinese radiology reports. After selecting the optimal full-parameter teacher model and generating mass soft labels, the tuned lightweight student model reaches 0.90 average precision, 0.49 recall and 0.62 F1-score, with inference speed boosted more than 40-fold. This study offers an innovative, lightweight and expandable technical scheme for the intelligent quality control of Chinese radiology reports.

Materials and methods

The dataset for this study was sourced from the Anhui Province Medical Imaging Cloud Platform. All included records are raw radiology reports submitted by primary-level hospitals without prior manual professional review, comprising 118,717 de-identified Chinese radiology reports that cover X-ray, CT, and MRI modalities. All private patient information was completely erased in strict compliance with medical ethics and data security management regulations. This study utilized de-identified data and open-source LLMs. The study was approved by the Ethics Committee of the First Affiliated Hospital of Anhui University of Chinese Medicine, which permitted the use of de-sensitized patient datasets for retrospective research and waived the requirement for written informed consent. All methods were performed in accordance with the relevant guidelines and regulations. The overall technical workflow of this study is illustrated in Fig. 2.

Quality Control Rule Formulation

This study rigorously developed a standardized QC rule system comprising 4 major categories and 13 quantifiable, detectable, and evaluable rules, strictly based on the Management and QC Standards for Radiology Departments (2017 Edition) issued by the Chinese Medical Association Radiology Branch, and in conjunction with clinical practice in Chinese radiology report writing. The specific rules are detailed in Table 1.

Table 1
Standardized quality control rule system for Chinese radiology reports

Category	Quality Control Rule
Report Logical Consistency	No internal contradictions within the imaging findings description
	No contradictions or anatomical discrepancies between imaging findings and impressions
	No confusion between patient gender and anatomical feature descriptions
Report Content Completeness	Contrast enhancement patterns must be described for enhanced CT/MR lesions
	Differential diagnosis must be provided for indeterminate lesions
	Anatomical location must be supplemented when only postoperative follow-up is mentioned
	Prostate Imaging Reporting and Data System(PI-RADS) classification must be performed for MR prostate positive lesions
Lesion Description Standardization	Breast Imaging Reporting and Data System(BI-RADS) classification must be performed for X-ray mammography positive lesions
	Measurable lesions must be explicitly annotated with size
	Lesions requiring description must be annotated with density/signal intensity
	Disease diagnostic terms must not be used in the imaging findings section
Textual and Format Standardization	Non-standard abbreviations, misspellings, and repeated punctuation marks must not be used
	Impressions must be ordered by clinical significance, from most to least important

Prompt Engineering Design for Natural Data Quality Control

General LLMs generate judgment errors under complex radiology QC rules, caused by insufficient medical knowledge, biased understanding of specifications and long-text information loss¹⁸. Traditional prompting loads all QC-rules, keyword screening and medical semantic reasoning into one lengthy prompt, raising model burden, obscuring rule boundaries, impairing keyword recognition and triggering extra false-positive and false-negative outcomes¹⁹. To address this bottleneck, we replaced the crude

“full-rule one-shot input” strategy with a fine-grained rule-decomposition, code–model hybrid prompt framework. Each of the 13 QC rules was divided into bounded and unbounded subtasks. Bounded subtasks involve explicit quantitative criteria or enumerable keywords and are handled by executable scripts through string matching, numerical validation, and list lookup, reducing inference overhead and avoiding LLM errors in rigid structured checks²⁰. Unbounded subtasks require cross-sentence consistency assessment, semantic contradiction detection, and medical reasoning; these are handled by the teacher LLM using prompts with standardized constraints and demonstration cases. This division removes mechanical validation from the LLM, reduces redundant input and long-context interference, and mitigates rule-interpretation bias and information loss, thereby reducing QC misjudgments and omissions. The rule-specific decomposition is summarized in Table 2.

Table 2
Fine-grained task decomposition of quality control rules

Rule Number	Rule Description	Bounded Subtasks (Code-Based)	Unbounded Subtasks (Model-Based)
1	Lesion Size Measurement	Morphological patterns (nodular, mass-like, round, etc.) with subsequent cm/mm units	Measurement determination for semantically flexible scenarios (e.g., gallbladder enlargement, prostatic enlargement)
2	Contrast Enhancement Description Required	Determination of CT/MR contrast-enhanced scan, exclusion of plain scan, Computed Tomography Angiography(CTA),Magnetic Resonance Angiography(MRA), angiography, and perfusion imaging	Semantic comprehension to assess whether lesion enhancement is described
3	Internal Contradictions in Imaging Findings	—	Comprehension of semantic conflicts and identification of contextual contradictions
4	Contradictions Between Findings and Impressions	—	Cross-paragraph assessment of anatomical location, description, and conclusion consistency
5	Differential Diagnosis for Indeterminate Lesions	—	Identification of indeterminate descriptions and assessment of whether differential diagnosis is provided
6	Textual and Format Standardization	—	Recognition of misspellings, grammatical errors, and non-standard expressions
7	Postoperative Follow-up Description Required	—	Assessment of whether effective description is provided for postoperative anatomical sites
8/9	PI-RADS/BI-RADS Classification	Modality and anatomical matching, determination of whether lesion classification is absent	Assessment of whether corresponding classification description is completed
10	Impression Ordering by Clinical Significance	Enumeration and sequence judgment of significant lesions	Semantic parsing of impressions with non-uniform formatting
11	Density/Signal Intensity Description	Modality, lesion morphology, density/signal lexicon, and angiography matching determination	Semantic determination of density description exemption under contrast-enhanced examination
12	Prohibited Disease Terms in Imaging Findings	Semantic unit segmentation combined with character-level retrieval from a disease terminology list	—
13	Gender-Anatomical Feature Conflict	Gender field and heterosexually anatomical taboo term retrieval	—

Based on the fine-grained decomposition, this study constructed a hierarchical prompt framework comprising four components: (1)Task Boundary Declaration: Explicitly stating that bounded subtasks have been completed by code, with the model handling only unbounded semantic reasoning.(2)Chain-of-Thought Reasoning Guidance: Requiring the model to sequentially extract modality, anatomical location, lesion characteristics, and impressions, followed by stepwise rule-based judgment, with output of contradictory sentence pairs and determination rationale.(3)Few-Shot Example Injection: Providing representative examples for each rule to standardize model judgment criteria.(4)Structured Output Constraints: Mandating output of violation type, violation location, contradictory sentence pairs, and semantic parsing results to facilitate subsequent soft-label generation.

Prompt Engineering Design for Artificial Data Generation

To address scarce positive samples and class imbalance in selected QC rules, we developed a standardized synthetic-data workflow that preserved radiology-report style and assigned explicit violation labels. Teacher-model-reviewed reports fully compliant with the target rule were selected as seeds. Rule-specific prompts directed the teacher LLM to introduce one predefined violation while retaining the anatomy, imaging modality, and core clinical information. Each synthetic report received a unique label. Random manual audits verified the intended error, absence of additional violations, and label accuracy. Meanwhile, samples failing these checks were discarded.

Teacher Model Selection and Evaluation

This study selected three full-parameter generative large language models as candidate teacher models: DeepSeek-v3²¹, Doubao-seed1.6-non-thinking, and Kimi-k2²². Dedicated prompts were designed for the QC task of standardized Chinese radiology reporting, with unified Application Programming Interface(API) call parameters. Precision, Recall, F1-score and average per-report inference time were adopted to evaluate teacher models.

Large-Scale Dataset Generation

The selected teacher model automatically conducted standardized QC on all 118,717 real clinical reports and produced annotations for 13 rules; a report violating a rule was labeled positive for that rule. To address scarcity and imbalance of low-frequency violations, dedicated prompts generated targeted synthetic positive samples to improve coverage of rare error patterns²³. Three radiologists randomly evaluated their clinical and medical-logical validity, reporting compliance, and training suitability. Validated synthetic samples were combined with natural reports to construct a balanced, reliable training set for student-model distillation and optimization.

Student Model Knowledge Distillation

This study selected three lightweight small-parameter models as student models: Qwen3-4B-Instruct-2507²⁴, SmoLLM3-3B, and DeepSeek-R1-Distill-Qwen-7B²⁵. Knowledge distillation training was performed using the Low-Rank Adaptation (LoRA) fine-tuning strategy²⁶ to transfer the standardized QC knowledge for radiology reports acquired by the teacher model to the lightweight student models.

Student Model Selection and Clinical Adaptability Assessment

On an independently human-annotated test set, average inference time per report was prioritized as the core metric, combined with Precision, Recall, and F1-score for comprehensive evaluation, to select a lightweight QC model with both high performance and low latency. Clinical testing was conducted on the Anhui Province Medical Imaging Cloud Platform, ultimately yielding a deployable radiology report standardization QC tool for clinical scenarios.

Test Set and evaluation metrics

An independent manually annotated test set covering all 13 QC rules was constructed from X-ray, CT, and MRI reports across head, neck, chest, abdomen, pelvis, extremities and urinary system. Three radiologists independently and blindly labeled candidate reports rule by rule for violation type. Labels were cross-checked, and discrepancies were jointly reviewed and resolved by experts to establish the final reference standard.

Model evaluation relied on Precision, Recall, F1-score and average per-report inference time. Standard binary classification definitions were applied: True Positive, False Positive, True Negative and False Negative count model prediction outcomes.

Precision calculates the ratio of genuine rule violations within all model-flagged cases, Recall quantifies the detection rate of real violations, and F1-score integrates the two metrics via their harmonic mean.

RESULTS

Test Set Sample Distribution

We manually constructed a test set comprising 734 radiology reports. Samples violating a specific rule were labeled as positive samples under that rule. The distribution of positive samples across the QC rules in the test set is illustrated in Fig. 3.

Teacher Model Selection Results

To select the optimal teacher model for QC of Chinese radiology report standardization, this study conducted a comparative evaluation of three mainstream full-parameter large language models: DeepSeek-v3, Doubao-seed1.6-non-thinking, and Kimi-k2. The comparison results of models and metrics are presented in Table 3. Doubao-seed1.6-non-thinking yields the best comprehensive performance and attains a precision of 0.94; accordingly, it is chosen as the teacher model.

Table 3
Metric Comparison of Candidate Teacher Large-Language-Models

	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	Rule 7	Rule 8	Rule 9	Rule 10	Rule 11	Rule 12	Rule 13	Average
Kimi-k2														
Precision	1	0.67	0.73	0.63	0.8	0.54	1	1	1	1	1	0.99	1	0.87
Recall	0.71	0.55	0.73	0.32	0.7	0.8	0.53	0.96	1	0.94	0.24	0.85	0.58	0.68
F1	0.83	0.6	0.73	0.42	0.74	0.65	0.69	0.98	1	0.97	0.38	0.92	0.74	0.74
DeepSeek-v3														
Precision	1	0.66	0.84	0.74	0.88	0.71	1	0.98	1	1	0.93	0.99	1	0.9
Recall	0.71	0.49	0.42	0.34	0.59	0.67	0.4	0.96	0.98	0.94	0.29	0.86	0.59	0.63
F1	0.83	0.56	0.56	0.47	0.71	0.69	0.57	0.97	0.99	0.97	0.44	0.92	0.74	0.72
Doubao-seed1.6-non-thinking														
Precision	1	0.74	0.93	0.76	0.97	0.97	0.9	1	1	1	0.97	0.99	1	0.94
Recall	0.71	0.4	0.2	0.29	0.45	0.42	0.45	0.96	0.98	0.94	0.31	0.86	0.59	0.58
F1	0.83	0.52	0.34	0.42	0.61	0.59	0.6	0.98	0.99	0.97	0.47	0.92	0.74	0.69

Distribution of Error Types in the Training Set Reports

The natural dataset for knowledge distillation contained 35,590 positive samples and 10,000 negative samples. The detailed distribution of positive labels across all quality control rules is listed in Electronic Supplementary Material Table S1.

Results of Artificial Sample Augmentation

Artificial sample expansion was carried out to alleviate insufficient positive samples and category imbalance for four low-frequency quality control rules. Supplementary manual quality inspection results and sample expansion volumes are available in Electronic Supplementary Material Table S2.

Performance of the Student Model After Knowledge Distillation

This study compared macro-averaged precision, recall and F1-score across three lightweight student models and Doubao-seed1.6-non-thinking on the test set, and the outcomes are shown in Fig. 4. Qwen3-4B-Instruct-2507 obtained the optimal

comprehensive performance at a precision of 0.90, which exceeded DeepSeek-R1-Distill-Qwen-7B and SmolLM3-3B.

Inference time of the optimal teacher model and three lightweight student models was compared on the test set, with results illustrated in Fig. 5. The Doubao-seed1.6-non-thinking teacher model took an average of 18.332 seconds per report. All student models achieved markedly higher inference efficiency. Qwen3-4B-Instruct-2507 yielded the shortest average inference time of 0.41 seconds, followed by SmolLM3-3B with a time of 0.513 seconds and DeepSeek-R1-Distill-Qwen-7B with a time of 0.719 seconds.

Real-World Clinical Validation Results

Qwen3-4B-Instruct-2507 was deployed on a clinical imaging cloud platform and processed 16,779 real radiology reports over one week. Attending radiologists revised 792 reports, and 587 revisions, 74 percent of all modified records, were completed according to quality control alerts generated by this model. Report modification statistics for each standard rule are listed in Table 4.

Table 4
Quality Control Performance in Real Clinical Environment

Rule	Accepted Revisions	Unaccepted Revisions	Total Revisions	Acceptance Rate
Rule 1	80	74	154	0.52
Rule 3	28	4	32	0.88
Rule 4	74	27	101	0.73
Rule 5	55	45	100	0.55
Rule 6	203	18	221	0.92
Rule 7	2	0	2	1
Rule 11	7	7	14	0.5
Rule 12	133	30	163	0.82
Rule 13	5	0	5	1
Total	587	205	792	0.74

Evaluation yielded Precision values exceeding 0.7 for Rule 3,4,6,7,12,13. Precision of Rule 1 decreased from 0.98 to 0.52, Precision of Rule 5 decreased from 0.91 to 0.55, and Precision of Rule 11 decreased from 0.77 to 0.50, with an overall Precision reduction of 0.16. The model achieved an average inference time of 0.28 seconds per report.

DISCUSSION

Key Findings

First, precision served as the major screening metric to decrease hazardous false-positive results. Doubao-seed1.6-non-thinking was selected as the optimal teacher model with an average precision of 0.94, ahead of DeepSeek-v3 at 0.90 and Kimi-k2 at 0.87.

Second, a synthetic-data pipeline created 1261 positive samples to tackle insufficient positive instances for infrequent quality-control rules. Random inspection produced an overall qualification rate of 0.958. The integrated dataset realized balanced category distribution for distillation-oriented training.

Finally, knowledge distillation transferred radiology-report quality-control knowledge into lightweight student models. Qwen3-4B-Instruct-2507 achieved the optimum average precision of 0.90, merely 0.04 lower than the teacher model, and delivered

inference speed improved by more than 40-fold. The hybrid prompt framework separated quality-control tasks for stable soft-label generation and supplies a reference template for medical-oriented prompt design.

Comparison with Previous Studies

Gertz et al.⁸ used cloud-hosted GPT-4 API to evaluate only 200 English reports with five error types, introducing persistent cost and data privacy risks. Sun et al.⁹ detected four error classes in English reports using mostly synthetic data, which required a 70B model deployed on two Graphics Processing Units (GPUs). Yan et al.¹⁰ adopted commercial cloud APIs for six-category error screening limited to ultrasound data, also suffering privacy and long-term operational cost issues. In contrast, this study built a 13-rule QC framework using 118,717 real Chinese X-ray, CT and MRI reports. The 4B lightweight model, optimized via knowledge distillation, runs locally on a single L40S GPU and eliminates reliance on cloud services.

Limitations of the study

First, this study performs standardized quality-control for Chinese radiology reports from logical consistency, content integrity, lesion-description, text and format specifications. Diagnostic-accuracy inspection and image-involved multimodal quality-control are excluded, and report-image consistency is not assessed.

Second, artificially constructed samples have limited error-type diversity and cannot cover all violation cases of each quality-control rule.

Third, model recall remains to be optimized, especially for open-ended rules including conflicts between imaging observations and conclusions and differential-diagnosis inspection of uncertain lesions.

Fourth, the model only identifies wrong sentences without revision guidance. Fixed prompt settings and varied writing habits among radiologists restrict subsequent report modification.

Clinical Applicability and Value

This lightweight model delivers two core clinical values. First, it automates large-scale QC for X-ray, CT and MRI reports to detect four types of textual nonstandardization, replacing inconsistent, labor-intensive manual inspection and reducing subjective bias and clinical risks, so radiologists can focus on diagnosis rather than repetitive review work. Second, its small parameter size and fast inference support local deployment on regular hospital servers without expensive hardware, cutting deployment costs and avoiding patient data leakage risks from cloud services, which fits the demand of primary hospitals short of QC staff and high-end computing resources.

Directions for Future Research

First, QC coverage shall be broadened to break the restriction of text-only standardization-oriented inspection. A dual-mode system combining standard-check and diagnostic-accuracy evaluation can be built, and matched images can be introduced for multimodal text-image joint quality-control.

Second, subsequent optimization targets model capability, distillation strategy and candidate-model screening. Sufficient clinical samples and chain-of-thought prompts will enhance the teacher-model reasoning for complicated semantic rules. Attention-aware and adversarial distillation methods will lift student-model recall and narrow the performance gap against the teacher model. Additional lightweight models will be evaluated for better comparative generality.

Moreover, wider application and technical integration are required to promote research value. The lightweight QC model can be adapted for pathological reports and medical progress notes, so as to develop a universal inspection framework for diverse clinical texts.

Conclusion

This method yields a lightweight local-deployable LLM dedicated to standardized QC of Chinese radiology reports. Despite a mild precision drop in real clinical scenarios, the model delivers over forty-fold faster inference while eliminating cloud data privacy risks. Balanced detection precision, efficient inference and clinical acceptance render this framework a feasible tool for routine radiology report quality management at medical facilities.

Abbreviations

PI-RADS Prostate Imaging Reporting and Data System

BI-RADS Breast Imaging Reporting and Data System

CT Computed Tomography

MRI Magnetic Resonance Imaging

CTA Computed Tomography Angiography

MRA Magnetic Resonance Angiography

LLM Large Language Model

LoRA Low-Rank Adaptation

API Application Programming Interface

GPU Graphics Processing Unit

TP True Positive

FP False Positive

TN True Negative

FN False Negative

P Precision

R Recall

QC Quality Control

Declarations

Data availability

The institutional data used in this study are not publicly available due to compliance with patient privacy protection but are available from the corresponding author on reasonable request.

Code availability

The code generated during this study are available in the ScienceDB data <https://doi.org/10.57760/sciencedb.42285> .

ACKNOWLEDGMENTS

This study was supported by The University Synergy Innovation Program of Anhui Province under Grant No. GXXT-2022-031.

Author contributions

Jiang Duoning, conceptualization, methodology, software, investigation, validation, visualization, writing – original draft, writing – review & editing; Han Xiao, validation, data curation; Xu Ke validation, data curation; Li Chuanfu, funding acquisition, resources, supervision, project administration. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Chuanfu Li.

References

1. Salam, B. et al. Large language models for error detection in radiology reports: a comparative analysis between closed-source and privacy-compliant open-source models. *Eur. Radiol.* **35**, 4549–4557. <https://doi.org/10.1007/s00330-025-11438-y> (2025).
2. Toker, I., Jansen, S., Knoche, J. & Lorenz, D. Radiology reporting in the age of artificial intelligence: implications for patient safety. *BMJ Qual. Saf.* <https://doi.org/10.1136/bmjqs-2025-018928> (2026).
3. Vosschenrich, J. et al. Revealing the most common reporting errors through data mining of the report proofreading process. *Eur. Radiol.* **31**, 2115–2125. <https://doi.org/10.1007/s00330-020-07306-6> (2021).
4. Pan, F. et al. RadCLARE: an automated clinical language engine for detecting semantic errors in radiology reports. *Eur. Radiol. Exp.* **9**, 120. <https://doi.org/10.1186/s41747-025-00659-x> (2025).
5. Wang, K. et al. in *Proceedings of the 34th ACM International Conference on Information and Knowledge Management* 3020–3030 (2025).
6. Fei, X. et al. Quality Management of Pulmonary Nodule Radiology Reports Based on Natural Language Processing. *Bioeng. (Basel)*. **9**. <https://doi.org/10.3390/bioengineering9060244> (2022).
7. Li, J., Zhou, Z., Lyu, H. & Wang, Z. Large language models-powered clinical decision support: enhancing or replacing human expertise? *Intell. Med.* **5**, 1–4. <https://doi.org/10.1016/j.imed.2025.01.001> (2025).
8. Gertz, R. J. et al. Potential of GPT-4 for Detecting Errors in Radiology Reports: Implications for Reporting Accuracy. *Radiology* **311**, e232714. <https://doi.org/10.1148/radiol.232714> (2024).
9. Sun, C. et al. Generative Large Language Models Trained for Detecting Errors in Radiology Reports. *Radiology* **315**, e242575. <https://doi.org/10.1148/radiol.242575> (2025).
10. Yan, Y. et al. The use of large language models in detecting Chinese ultrasound report errors. *NPJ Digit. Med.* **8**, 66. <https://doi.org/10.1038/s41746-025-01468-7> (2025).
11. Dennstadt, F., Hastings, J., Putora, P. M., Schmerder, M. & Cihoric, N. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *NPJ Digit. Med.* **8**, 143. <https://doi.org/10.1038/s41746-025-01476-7> (2025).
12. Yang, R. et al. Large language models in health care: Development, applications, and challenges. *Health Care Sci.* **2**, 255–263. <https://doi.org/10.1002/hcs2.61> (2023).
13. Nagarajan, R. et al. Economics and Equity of Large Language Models: Health Care Perspective. *J. Med. Internet Res.* **26**, e64226. <https://doi.org/10.2196/64226> (2024).
14. Garg, M., Raza, S., Rayana, S., Liu, X. & Sohn, S. The Rise of Small Language Models in Healthcare: A Comprehensive Survey. *Comput. Sci. Rev.* **62** <https://doi.org/10.1016/j.cosrev.2026.100999> (2026).
15. Kang, M., Lee, S., Baek, J., Kawaguchi, K. & Hwang, S. J. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *Adv. Neural. Inf. Process. Syst.* **36**, 48573–48602 (2023).
16. Woo, E. G., Burkhart, M. C., Alsentzer, E. & Beaulieu-Jones, B. K. Synthetic data distillation enables the extraction of clinical information at scale. *NPJ Digit. Med.* **8**, 267. <https://doi.org/10.1038/s41746-025-01681-4> (2025).

17. Suleman, M. U. et al. Assessing the generalizability of artificial intelligence in radiology: a systematic review of performance across different clinical settings. *Ann. Med. Surg. (Lond)*. **87**, 8803–8811. <https://doi.org/10.1097/MS9.0000000000004166> (2025).
18. Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H. & Luo, Y. A comparative study of pretrained language models for long clinical text. *J. Am. Med. Inf. Assoc.* **30**, 340–347. <https://doi.org/10.1093/jamia/ocac225> (2023).
19. Colangelo, M. T., Guizzardi, S., Meleti, M., Calciolari, E. & Galli, C. How to Write Effective Prompts for Screening Biomedical Literature Using Large Language Models. *BioMedInformatics* **5** (2025). <https://doi.org/10.3390/biomedinformatics5010015>
20. Kim, J. et al. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Sci. Rep.* **15**, 39426. <https://doi.org/10.1038/s41598-025-22940-0> (2025).
21. Liu, A. et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).
22. Team, K. et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534* (2025).
23. Ashofteh Barabadi, M., Zhu, X., Chan, W. Y., Simpson, A. L. & Do, R. K. G. Targeted generative data augmentation for automatic metastases detection from free-text radiology reports. *Front. Artif. Intell.* **8**, 1513674. <https://doi.org/10.3389/frai.2025.1513674> (2025).
24. Yang, A. et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).
25. Guo, D. et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).
26. Hu, E. J. et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

Figures

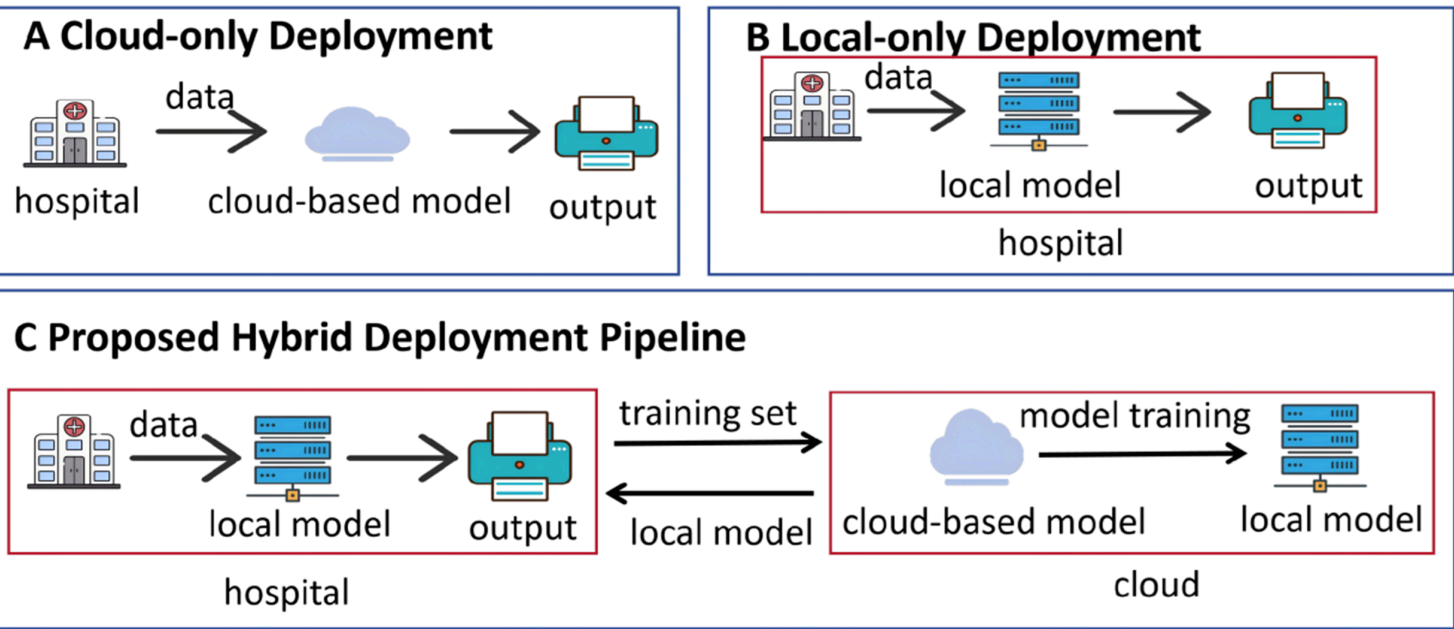


Figure 1

Comparison of three deployment pipelines for radiology report intelligent QC. (A) Cloud-only Deployment. Raw radiology-report data are uploaded to third-party cloud platforms and processed by closed-source LLMs. (B) Local-only Deployment. All model training and inference run on hospital-local servers without external data transfer. (C) The Hybrid Deployment proposed in this work. Only de-identified training data are uploaded to the cloud for full-parameter teacher-model training and soft-label generation, and the distilled lightweight student LLM is deployed on in-hospital local devices to balance clinical-data privacy and error-detection accuracy.

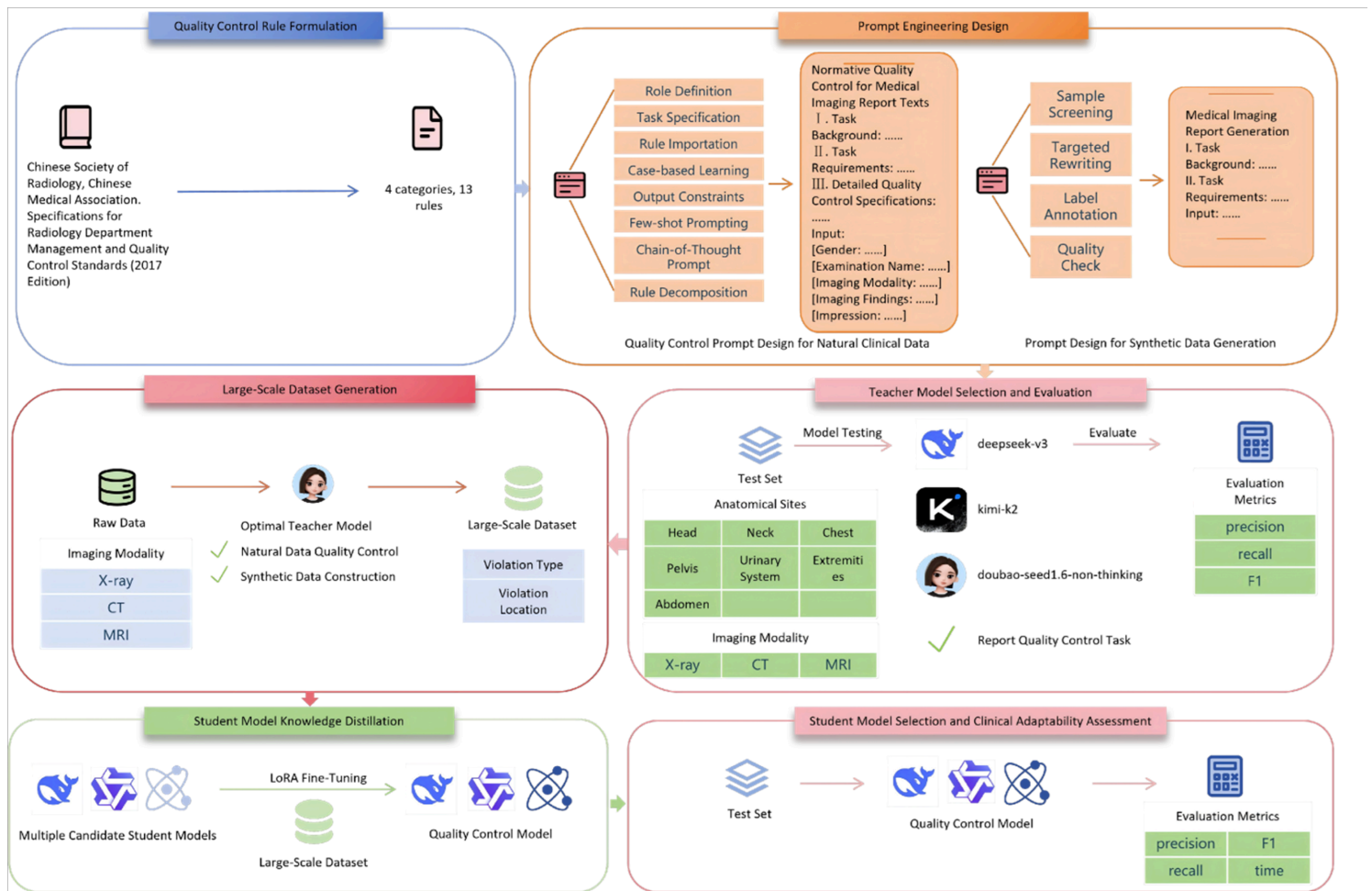


Figure 2

Technical workflow for standardized QC of Chinese radiology reports using knowledge distillation. The pipeline includes six stages: (1) QC rule formulation; (2) prompt design for quality inspection and data augmentation; (3) teacher LLM selection and benchmarking; (4) balanced dataset construction using real and synthetic data; (5) knowledge distillation with low-rank adaptation for lightweight student models; and (6) student model screening and clinical validation based on inference efficiency and overall performance to identify the optimal model for deployment.

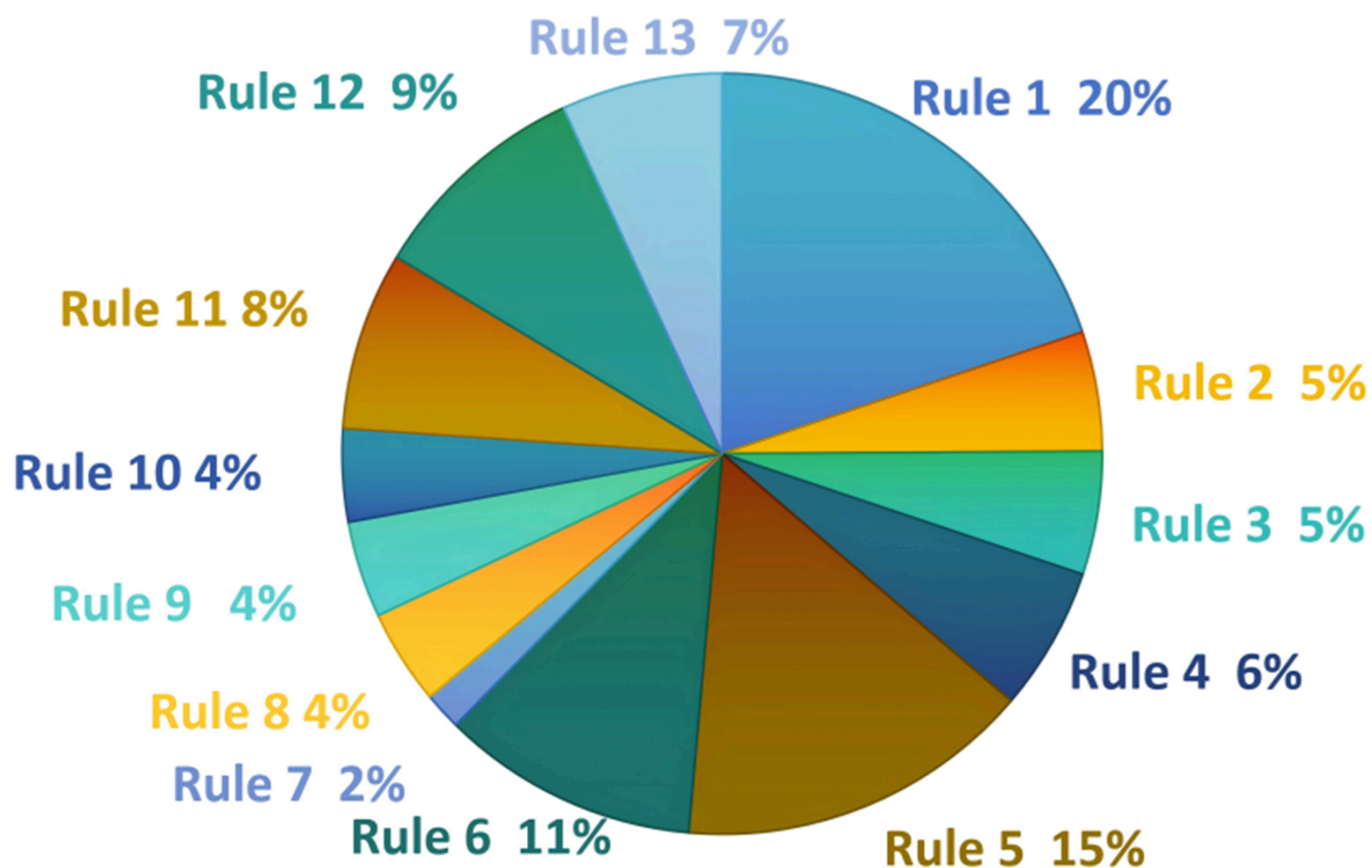


Figure 3

Distribution of positive samples across QC rules in the test set. Rule 1 has the maximum positive ratio of 20%, followed by Rule 5 (15%), Rule 6 (11%) and Rule 12 (9%). Rule 7 features the minimum ratio of 2%, and the rest rules hold positive proportions between 4% and 8%.

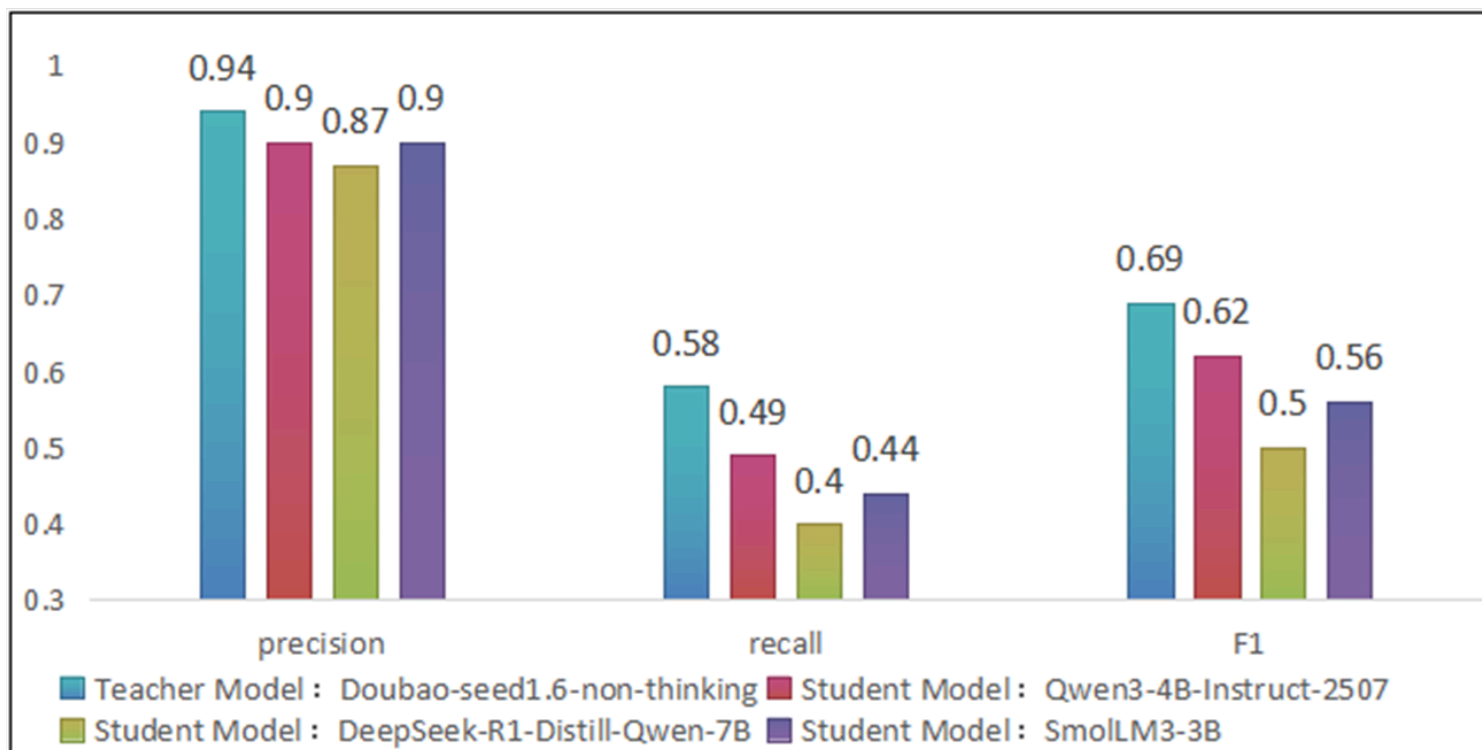


Figure 4

Performance comparison between student models and the optimal teacher model

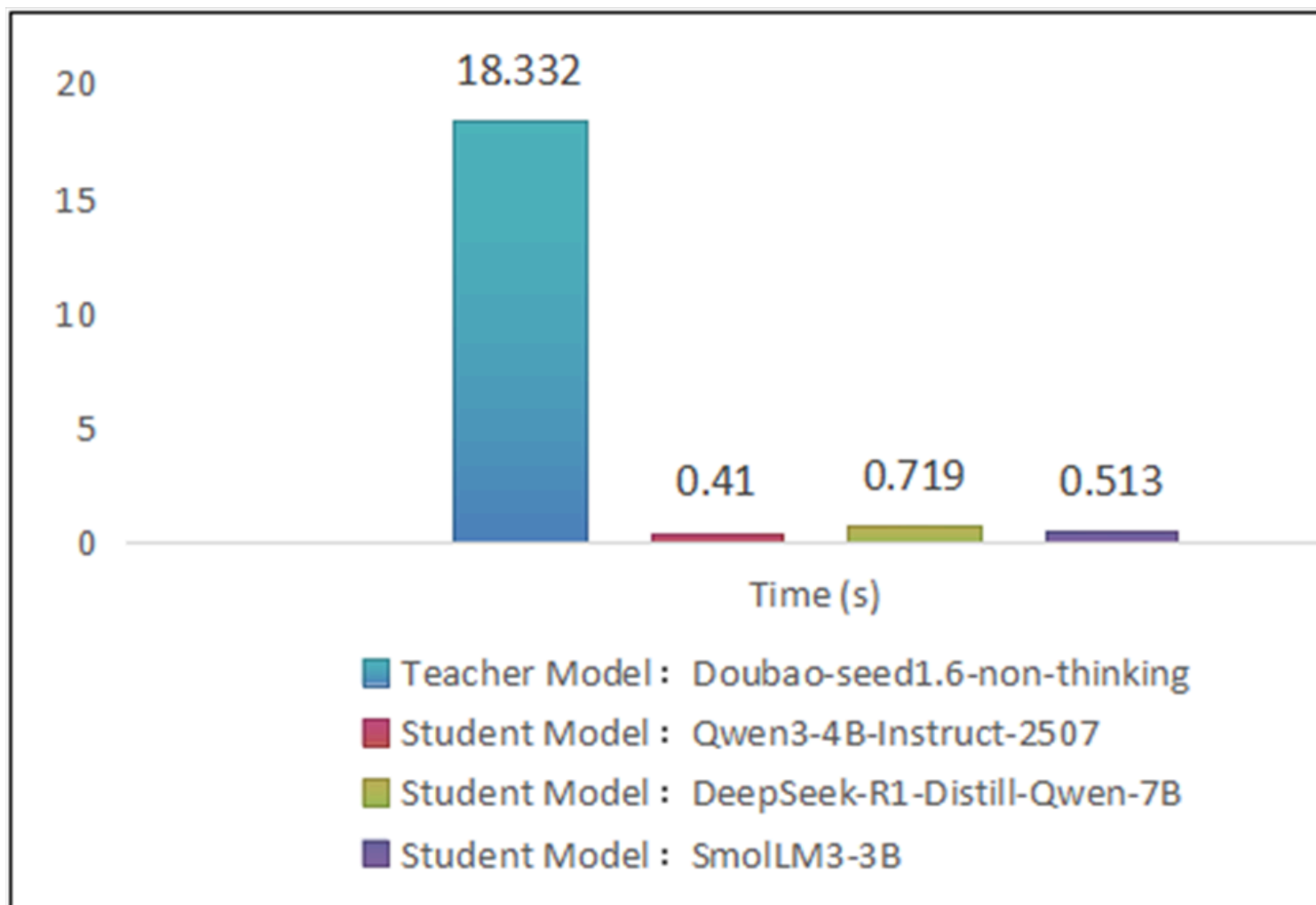


Figure 5

Comparison of inference time between student and teacher models

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ElectronicSupplementaryMaterial.docx](#)