Applied Artificial Intelligence Research

INNOVIAIR

Chinese Large Language Models for Patient-Friendly Knee MRI Summaries in Sports Injury: A Guideline-Anchored Study of Quality and Robustness

Shangan Zhou[a,*], Yiming Wang[b]

[a]College of Physics and Electronic Information Engineering, Zhejiang Normal University, Jinhua, Zhejiang 321004, China

[b]College of Life Sciences, Zhejiang Normal University, Jinhua, Zhejiang 321004, China

ARTICLEINFO

Keywords:
Large Language Models
Knee Magnetic Resonance Imaging
Sports Injury
Plain-Language Reporting
Patient Communication
Readability
Semantic Fidelity
Robustness

ABSTRACT

Reports describing sports-related knee MRI findings are usually written for clinicians and can be difficult for patients to interpret. We compared the quality and consistency of five widely used Chinese large language models (LLMs) when converting such reports into plain language. A guideline-anchored set of 50 simulated knee MRI reports was prepared, and the five models were tested with zero-shot prompting; DeepSeek-V4 was also tested with a few-shot prompt. The six experimental arms produced 300 summaries through the vendors' official web interfaces. Two blinded physicians rated factual accuracy (0-3), information coverage (0-5), and patient-oriented clarity (0-5). We additionally calculated mean readability grade level (mRGL) and BERT-based semantic similarity. Accuracy remained high across arms (2.60-2.83) without a significant pairwise difference after correction. Few-shot DeepSeek-V4 obtained the best coverage score (4.33 +/- 0.61), Doubao-Seed-2.0 generated the easiest text to read (mRGL 7.15), and GLM-5.2 remained closest to the wording of the source reports (similarity 0.72). The models were less alike in difficult combined injuries: Doubao-Seed-2.0 and Qwen3.8-Max showed the clearest decline, while few-shot DeepSeek-V4 had the lowest accuracy CV (12.7%). The findings indicate that Chinese flagship LLMs can produce generally accurate patient-facing knee MRI explanations, but they occupy different positions with respect to coverage, readability, fidelity, and robustness. Safe selection therefore requires more than a single average score.

1.Introduction

Knee trauma involving the anterior cruciate ligament (ACL) or the meniscus is a frequent reason for young and physically active people to seek orthopaedic assessment[1-2]. MRI is the principal non-invasive examination used to characterize these lesions and often becomes part of the discussion about observation, rehabilitation, or surgery[3-4]. The report, however, is normally composed for the referring clinician. It may contain expressions such as pivot-shift bone marrow oedema, bucket-handle tear, and ramp lesion, all of which are meaningful to specialists but opaque to many patients. Work on health literacy recommends that patient materials stay near a sixth-grade to eighth-grade reading level[5-6], whereas conventional imaging reports rarely meet that standard.

This communication gap is more consequential now that patients can often open their imaging results electronically before a clinical consultation. Policies that encourage rapid release of health information, including the information-blocking provisions of the US 21st Century Cures Act, have made direct access routine[7]. An unexplained abnormality can consequently be interpreted as a severe diagnosis, even when the finding is minor or requires clinical context. A plain-language companion that is accurate, complete, and appropriately cautious has therefore become part of patient-centred musculoskeletal communication rather than an optional convenience[5-6,8].

LLMs offer one possible way to produce that companion text at scale. Early investigations found that general chatbots could restate radiology and pathology reports with encouraging factual performance[9-10]. Later work introduced structured expert rubrics, comparison with physician responses, readability targets, and automated language measures[9-12]. Together, these studies provide a reproducible starting point, but the evidence base remains concentrated on English reports and predominantly Western systems. Orthopaedic sports medicine, despite involving patients who

* Corresponding author.
E-mail address: 3217432128@qq.com.
https://doi.org/10.65455/2c4jes43
Received 20 August 2026; Received in revised form 22 August 2026; Accepted 24 August 2026; Available 1 September 2026

are often highly motivated to understand their diagnosis, has been comparatively underexamined.

Knee MRI is a demanding setting in which to test patient-oriented rewriting. A useful summary must keep track of abbreviated structures such as ACL, PCL, and MCL; injury grade; the shape and location of a tear; spatial descriptors such as the posterior horn or femoral footprint; and secondary signs such as bone bruising or anterior tibial translation[3-4,13]. Several abnormalities may occur in the same knee. Omitting one of these elements can change the clinical meaning even if the remaining sentences are fluent. The task is therefore not simple translation: it is controlled compression that must retain the distinctions relevant to management under current guidance[13-15].

Chinese model families have developed rapidly. DeepSeek-V3 reported performance close to leading proprietary systems while using substantially fewer resources[16], and the GLM, Qwen, Kimi, and Doubao families have continued to release new flagship versions[17-19]. Their consumer web portals are readily available to a large Chinese-speaking population. Although representative survey data are not yet available, patients do in practice use public AI services to help interpret medical documents.

Evidence is still limited on how dependable these systems are when they simplify musculoskeletal imaging, especially when the report contains several injuries. Average performance can obscure a model's behaviour on the more difficult cases. This matters clinically because the patients most likely to receive dense, multi-lesion reports may also be the patients for whom an omitted grade, structure, or associated finding is most consequential.

Here, stability means that output quality remains reasonably consistent when the underlying reports differ in complexity and when the prompting condition changes. We represented it in three ways: within-arm coefficient of variation (CV) for expert scores, the spread of mean scores across four injury categories, and the weakest category-specific result. This definition treats poor tail performance as a deployment concern rather than allowing a strong overall mean to conceal it[8].

We therefore compared five Chinese flagship LLMs on a bank of 50 guideline-anchored simulated knee MRI reports under zero-shot prompting, with an additional few-shot arm for DeepSeek-V4. Human raters assessed accuracy, coverage, and clarity; automated measures captured readability and semantic proximity to the source. The prespecified aims were to (1) compare the models across these dimensions, (2) estimate the effect of one exemplar pair, and (3) examine whether quality changed across injury categories in a way that would affect patient-facing use.

2.Methods

2.1.Study design and simulated report bank

We performed a cross-sectional comparative evaluation from 10 to 16 August 2026. All cases were simulated and contained no identifiable patient information; consequently, institutional review and informed consent were not applicable. The report structure and three-part expert rubric were adapted from previous work on LLM-assisted medical simplification[9-10,12].

The stimulus bank contained 50 English knee MRI reports. Each record included a brief clinical history, detailed imaging findings, and a radiological impression (Supplementary File S1). The cases were divided into four clinically relevant groups: ACL injuries (group A, n = 15; KN-A01 to KN-A15), meniscal injuries (group B, n = 15; KN-B01 to KN-B15), combined ACL-meniscal injuries (group C, n = 10; KN-C01 to KN-C10), and other common sports injuries (group D, n = 10; KN-D01 to KN-D10). The ACL group included complete and partial tears, chronic change, mucoid degeneration, ganglion cysts, associated MCL injury, and tibial eminence avulsion. The meniscal group covered longitudinal vertical, bucket-handle, horizontal, radial, and root tears as well as discoid meniscus. The combined group included ramp lesions, lateral root tears, the unhappy triad, and chondral injury. The final group included PCL and MCL lesions, posterolateral corner injury, patellar dislocation, osteochondritis dissecans, and graft re-rupture.

The injury patterns and their embedded management implications were checked against the AAOS ACL guidance[3,14], the 2019 ESSKA consensus on traumatic meniscal tears[13], the 2016 ESSKA consensus on degenerative meniscal lesions[15], the Panther Symposium return-to-sport consensus[20], and the ACR criteria for acute knee trauma[4]. For each case, we specified a set of gold-standard points that a faithful patient explanation had to preserve; these points formed the reference for the accuracy assessment.

Two investigators trained in sports medicine drafted the cases through several rounds of review. Histories recorded age, sex, injured side, mechanism, and symptom duration. Findings described signal change, structural continuity, tear morphology, associated lesions, and joint effusion using conventional radiological terminology; the impression followed a standard clinical-report format. Each case was also classified as Single, Combined, or Special. The Special label covered atypical patterns such as mucoid degeneration, ganglion cyst, discoid meniscus, and tibial eminence avulsion in a skeletally immature patient. The investigators cross-checked every case against the source recommendations, including the management relevance of a longitudinal meniscal tear greater than 10 mm and a grade II associated MCL injury[13-15,20]. The full stimulus set, prompts, and gold-standard points are supplied in Supplementary File S1.

2.2.Models and prompting strategies

Six arms were included (Table 1). Five vendor models were evaluated with zero-shot prompting: DeepSeek-V4, Qwen3.8-Max, Kimi-K3, Doubao-Seed-2.0, and GLM-5.2. DeepSeek-V4 was evaluated a second time with a few-shot prompt containing one example pair. The systems were accessed through the vendors' official web portals between 10 and 16 August 2026, using default generation settings. A fresh conversation was opened for every case to avoid carry-over from earlier queries.

The design was intended to resemble ordinary patient use. We used web portals rather than APIs, retained default

sampling settings, and isolated each report in a new session. Each response was capped at 300 words, both to limit unbounded elaboration and to approximate a practical patient handout. Because web-hosted models can change without notice, data collection was completed within one week and the access interval was recorded as the version snapshot[16-19].

The zero-shot instruction asked the model to explain the report as a sports-medicine clinician speaking to a reader without medical training. It required preservation of the injured structure, severity or grade, tear morphology and location, and associated findings; requested plain language at approximately a sixth- to eighth-grade level; allowed short explanations of unavoidable terms; and prohibited new diagnoses or specific treatment advice. Responses were written in English and limited to 300 words. The few-shot condition placed one expert-written report-summary pair before the target report. Complete prompts are provided in Supplementary File S1.

Table 1. Experimental arms included in the comparison

Model arm	Developer	Access portal	Prompting	Outputs (n)
DeepSeek-V4 (Zero-Shot)	DeepSeek	chat.deepseek.com	Zero-Shot	50
Qwen3.8-Max (Zero-Shot)	Alibaba	tongyi.aliyun.com	Zero-Shot	50
Kimi-K3 (Zero-Shot)	Moonshot AI	kimi.com	Zero-Shot	50
Doubao-Seed-2.0 (Zero-Shot)	ByteDance	doubao.com	Zero-Shot	50
GLM-5.2 (Zero-Shot)	Zhipu AI	chatglm.cn	Zero-Shot	50
DeepSeek-V4 (Few-Shot)	DeepSeek	chat.deepseek.com	Few-Shot (1 exemplar pair)	50

Note: All six arms were queried through official web interfaces during 10-16 August 2026 with default generation settings and a new session for each item. Outputs were in English and could contain no more than 300 words.

2.3.Human evaluation

Two sports-medicine physicians independently reviewed the 300 summaries. Model labels and output order were hidden, and the summaries were randomized. Accuracy was scored from 0 to 3 against the gold-standard points: 0 indicated inconsistent or potentially harmful content, 1 partial consistency, 2 high consistency with minor errors, and 3 complete consistency. Coverage and understandability were each scored from 0 to 5, ranging respectively from incomplete to comprehensive and from jargon-heavy to clear. The physicians consulted the reference points for accuracy, discussed disagreements to reach consensus, and the mean of the two initial ratings was used for analysis. Agreement was summarized with quadratic-weighted Cohen kappa and ICC(3,1), both with 95% confidence intervals; interpretation followed Landis and Koch and Koo and Li[21,22].

2.4.Automated metrics

Readability was summarized as mean readability grade level (mRGL), the average of the Gunning Fog, Flesch-Kincaid, Automated Readability Index, and Coleman-Liau scores[23-26]. A lower value represents easier reading, and a value around 8 was treated as the upper boundary suggested for patient material[5,6]. Semantic fidelity was estimated with cosine similarity between BERT embeddings of the generated summary and its source report, using bert-base-uncased[27]. A higher value indicates greater proximity in wording and structure.

The four readability formulas emphasize different surface properties. Fog and Flesch-Kincaid are influenced by sentence length and multisyllabic words, ARI uses character and sentence counts, and Coleman-Liau relies on character statistics rather than syllable counting[23-26]. Their average was used to reduce dependence on any one formula. The BERT measure was interpreted jointly with expert accuracy: a lower similarity can reflect a useful re-expression rather than an error, although it can also signal deviation from the source.

2.5.Statistical analysis

The analysis was exploratory. The bank size of 50 simulated cases was selected with reference to comparable LLM report-simplification studies because a conventional a priori power calculation was not appropriate for this design[9-10,12]. Continuous variables are reported as mean +/- SD, with normality assessed by the Shapiro-Wilk test. Since every case was processed in every arm, between-arm comparisons were paired at the item level and used paired t tests or Wilcoxon signed-rank tests as appropriate[28]. Benjamini-Hochberg adjustment controlled the false discovery rate across 15 pairwise comparisons per metric[29]. For stability, we calculated CV (SD/mean) for each expert-rated dimension and compared means across the four injury groups. Tests were two-sided, with FDR-adjusted $P < 0.05$ considered significant.

The fully crossed structure improved efficiency by controlling for case difficulty in paired comparisons. A post-hoc sensitivity calculation suggested that 50 paired observations would provide approximately 80% power to detect a standardized mean difference of about 0.4 at a two-sided alpha of 0.05. This level was considered reasonable for the moderate-to-large contrasts of practical interest. The statistical analyses used standard software; item-level scores, unadjusted and adjusted P values, and the complete analysis output are supplied in Supplementary File S2.

3.Results

3.1.Expert scores

All 300 generated summaries were retained for analysis. Accuracy means were high and tightly grouped (Figure 1, Table 2). Few-shot DeepSeek-V4 had the highest mean (2.83 +/- 0.36), followed by zero-shot DeepSeek-V4 (2.79 +/- 0.39),

GLM-5.2 (2.78 +/- 0.42), Kimi-K3 (2.76 +/- 0.44), Qwen3.8-Max (2.69 +/- 0.45), and Doubao-Seed-2.0 (2.60 +/- 0.52); none of the pairwise accuracy contrasts remained significant after FDR correction (all FDR P > 0.14). Coverage separated the arms more clearly. Few-shot DeepSeek-V4 scored 4.33 +/- 0.61, followed by Kimi-K3 at 4.12 +/- 0.52 and zero-shot DeepSeek-V4 at 4.11 +/- 0.46. Doubao-Seed-2.0 (3.60 +/- 0.60; FDR P = 0.0013) and GLM-5.2 (3.79 +/- 0.53; FDR P = 0.0281) were lower than zero-shot DeepSeek-V4. Doubao-Seed-2.0 had the best zero-shot understandability score (4.40 +/- 0.49), while GLM-5.2 had the lowest (3.98 +/- 0.39), significantly below zero-shot DeepSeek-V4 (4.28 +/- 0.49; FDR P = 0.014).

The accuracy distribution supported this ceiling effect. A perfect score of 3 was assigned to 58% of Doubao-Seed-2.0 outputs and to 80% of few-shot DeepSeek-V4 outputs; the other zero-shot arms fell between 66% and 76%. Scores of 2 or below occurred most often with Doubao-Seed-2.0 (32%) and Qwen3.8-Max (28%) and least often with few-shot DeepSeek-V4 (14%). No summary received an accuracy score of 0, and neither physician identified a fabricated injury or an invented treatment recommendation.

Table 2. Mean expert ratings by model arm, based on two blinded raters

Model arm	Accuracy (0-3)	Comprehensiveness (0-5)	Understandability (0-5)
DeepSeek-V4 (Zero-Shot)	2.79 ± 0.39	4.11 ± 0.46	4.28 ± 0.49
Qwen3.8-Max (Zero-Shot)	2.69 ± 0.45	3.93 ± 0.52	4.33 ± 0.51
Kimi-K3 (Zero-Shot)	2.76 ± 0.44	4.12 ± 0.52	4.31 ± 0.46
Doubao-Seed-2.0 (Zero-Shot)	2.60 ± 0.52	3.60 ± 0.60	4.40 ± 0.49
GLM-5.2 (Zero-Shot)	2.78 ± 0.42	3.79 ± 0.53	3.98 ± 0.39
DeepSeek-V4 (Few-Shot)	2.83 ± 0.36	4.33 ± 0.61	4.43 ± 0.54

Note: Across the 18 arm-by-dimension combinations, weighted kappa was 0.71-0.90 and ICC(3,1) was 0.71-0.91; per-arm estimates and 95% CIs are reported in Supplementary File S2.

Table 3. Pairwise expert-score comparisons with zero-shot DeepSeek-V4, using the Wilcoxon signed-rank test and Benjamini-Hochberg correction

Dimension	Comparison	Mean 1	Mean 2	FDR P	Significant
Accuracy	Qwen3.8-Max vs DeepSeek-V4 ZS	2.69	2.79	0.601	No
Accuracy	Kimi-K3 vs DeepSeek-V4 ZS	2.76	2.79	0.749	No
Accuracy	Doubao-Seed-2.0 vs DeepSeek-V4 ZS	2.60	2.79	0.143	No
Accuracy	GLM-5.2 vs DeepSeek-V4 ZS	2.78	2.79	0.884	No
Accuracy	DeepSeek-V4 FS vs DeepSeek-V4 ZS	2.83	2.79	0.612	No
Comprehensiveness	Qwen3.8-Max vs DeepSeek-V4 ZS	3.93	4.11	0.285	No
Comprehensiveness	Kimi-K3 vs DeepSeek-V4 ZS	4.12	4.11	0.929	No
Comprehensiveness	Doubao-Seed-2.0 vs DeepSeek-V4 ZS	3.60	4.11	0.0013	Yes
Comprehensiveness	GLM-5.2 vs DeepSeek-V4 ZS	3.79	4.11	0.028	Yes
Comprehensiveness	DeepSeek-V4 FS vs DeepSeek-V4 ZS	4.33	4.11	0.051	No
Understandability	Qwen3.8-Max vs DeepSeek-V4 ZS	4.33	4.28	0.749	No
Understandability	Kimi-K3 vs DeepSeek-V4 ZS	4.31	4.28	0.884	No
Understandability	Doubao-Seed-2.0 vs DeepSeek-V4 ZS	4.40	4.28	0.601	No
Understandability	GLM-5.2 vs DeepSeek-V4 ZS	3.98	4.28	0.014	Yes
Understandability	DeepSeek-V4 FS vs DeepSeek-V4 ZS	4.43	4.28	0.285	No

Note: ZS, zero-shot; FS, few-shot. Ten of the 15 mRGL contrasts remained significant after FDR adjustment, with Doubao-Seed-2.0 differing from every other arm. All 15 BERT-similarity contrasts were significant. Complete comparison matrices are in Supplementary File S2.

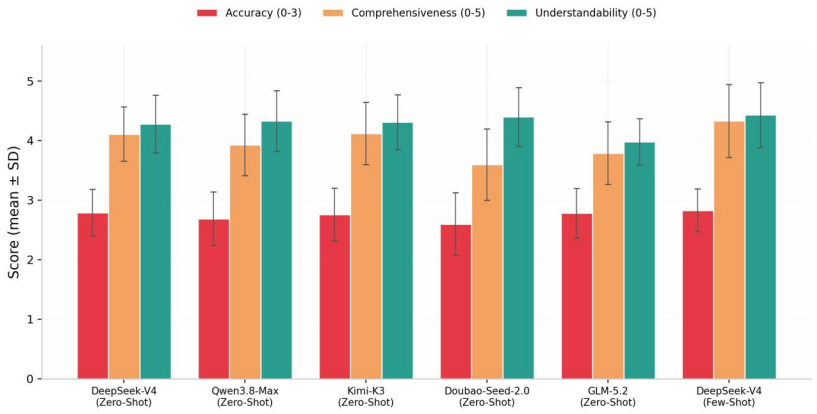

Fig 1. Mean expert ratings for accuracy, coverage, and understandability across six arms; error bars indicate SD over 50 cases

3.2.Readability

Readability varied more than accuracy (Figure 2, Table 4). Doubao-Seed-2.0 had the lowest mRGL (7.15 +/- 1.16), significantly below every other arm (all FDR P < 0.001), and was the only mean close to the eighth-grade reference point. Qwen3.8-Max (8.49 +/- 1.00) and zero-shot DeepSeek-V4 (8.64 +/- 1.19) were intermediate. Kimi-K3 (9.25 +/- 1.15), GLM-5.2 (9.04 +/- 1.16), and few-shot DeepSeek-V4 (9.46 +/- 1.11) were above a ninth-grade level. Adding an exemplar to DeepSeek-V4 significantly increased mRGL compared with its zero-shot condition (FDR P = 0.0009), despite the improvement in expert-rated coverage. Ten of 15 pairwise mRGL comparisons were significant after correction.

The threshold analysis showed the same ranking. At mRGL <= 8, 80% of Doubao-Seed-2.0 summaries met the target, compared with 30% for Qwen3.8-Max, 24% for zero-shot DeepSeek-V4, 20% for GLM-5.2, 12% for Kimi-K3, and 8% for few-shot DeepSeek-V4. Injury category contributed less than model choice; in most arms, cases in the other-injuries group produced slightly lower mRGL values than ACL or combined-injury cases.

Table 4. Readability grade level (mRGL) overall and within each injury group; smaller values indicate easier text

Model arm	Overall	A: ACL	B: Meniscus	C: Combined	D: Other
DeepSeek-V4 (Zero-Shot)	8.64 ± 1.19	8.58 ± 1.34	8.90 ± 1.20	8.78 ± 0.79	8.21 ± 1.28
Qwen3.8-Max (Zero-Shot)	8.49 ± 1.00	8.65 ± 0.90	8.46 ± 0.97	8.29 ± 1.38	8.48 ± 0.85
Kimi-K3 (Zero-Shot)	9.25 ± 1.15	9.51 ± 0.95	9.55 ± 0.92	9.15 ± 1.60	8.51 ± 1.03
Doubao-Seed-2.0 (Zero-Shot)	7.15 ± 1.16	7.51 ± 1.36	7.03 ± 0.84	7.14 ± 1.00	6.78 ± 1.41
GLM-5.2 (Zero-Shot)	9.04 ± 1.16	9.19 ± 1.46	8.55 ± 1.02	9.20 ± 0.78	9.38 ± 1.10
DeepSeek-V4 (Few-Shot)	9.46 ± 1.11	8.87 ± 0.93	9.49 ± 0.94	9.97 ± 0.92	9.80 ± 1.47

Note: mRGL is the mean of the Gunning Fog, Flesch-Kincaid, Automated Readability Index, and Coleman-Liau grade estimates.

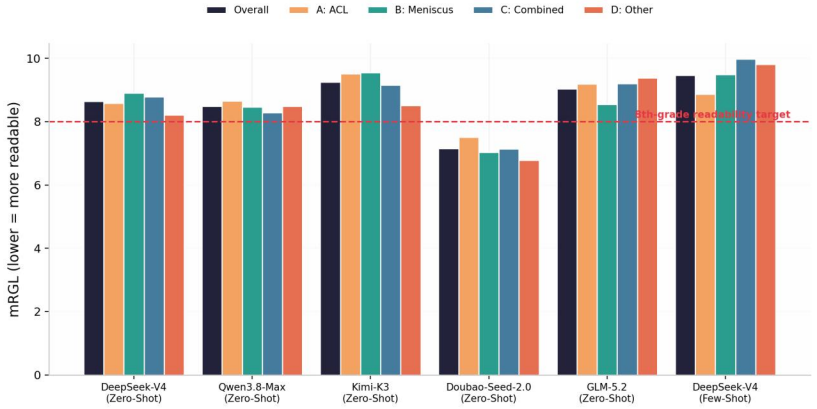

Fig 2. Overall and category-specific mRGL for the six arms

Note: A, ACL; B, meniscus; C, combined ACL and meniscal injury; D, other sports injuries. The dashed line marks the eighth-grade target for patient-directed material.

3.3.Semantic similarity to source reports

The models occupied clearly separated positions on BERT similarity (Figure 3, Table 5). GLM-5.2 was closest to the original reports (0.72 +/- 0.06), followed by Qwen3.8-Max (0.64 +/- 0.08), Kimi-K3 (0.60 +/- 0.09), zero-shot DeepSeek-V4 (0.56 +/- 0.08), Doubao-Seed-2.0 (0.51 +/- 0.07), and few-shot DeepSeek-V4 (0.48 +/- 0.07). Every pairwise contrast was significant after FDR adjustment. Thus, the arms ranged from conservative restatement (GLM-5.2) to more extensive restructuring (few-shot DeepSeek-V4).

GLM-5.2 combined the highest similarity with the smallest SD (0.06), suggesting a consistent tendency to stay close to the source. Few-shot DeepSeek-V4 also showed relatively low dispersion (SD 0.07), but around a much lower mean, implying that its more substantial rewriting was systematic. Kimi-K3 was the most variable (SD 0.09), moving between near-paraphrase and more reorganized explanations depending on the case.

Table 5. BERT semantic similarity between generated summaries and their source reports, overall and by injury category

Model arm	Overall	A: ACL	B: Meniscus	C: Combined	D: Other
DeepSeek-V4 (Zero-Shot)	0.56 ± 0.08	0.55 ± 0.10	0.59 ± 0.07	0.52 ± 0.08	0.54 ± 0.06
Qwen3.8-Max (Zero-Shot)	0.64 ± 0.08	0.64 ± 0.07	0.64 ± 0.07	0.67 ± 0.06	0.61 ± 0.12
Kimi-K3 (Zero-Shot)	0.60 ± 0.09	0.57 ± 0.09	0.60 ± 0.08	0.63 ± 0.10	0.61 ± 0.08
Doubao-Seed-2.0 (Zero-Shot)	0.51 ± 0.07	0.50 ± 0.08	0.52 ± 0.07	0.49 ± 0.05	0.52 ± 0.07
GLM-5.2 (Zero-Shot)	0.72 ± 0.06	0.71 ± 0.06	0.73 ± 0.05	0.74 ± 0.07	0.72 ± 0.08
DeepSeek-V4 (Few-Shot)	0.48 ± 0.07	0.51 ± 0.05	0.47 ± 0.08	0.47 ± 0.06	0.47 ± 0.07

Note: Similarity was calculated from bert-base-uncased embeddings for each source-summary pair.

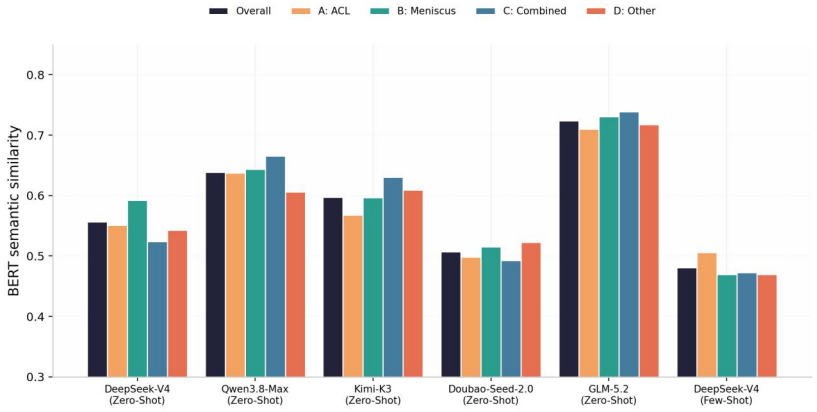

Fig 3. BERT similarity overall and by injury category

Note: Larger values indicate greater closeness to the wording and structure of the source report.

3.4.Cross-category stability

The stability profile was not uniform across models (Figure 4). Accuracy CV was lowest for few-shot DeepSeek-V4 (12.7%) and zero-shot DeepSeek-V4 (14.1%), followed by GLM-5.2 (15.1%), Kimi-K3 (16.1%), and Qwen3.8-Max (16.7%); Doubao-Seed-2.0 had the highest CV (20.2%). Combined ACL-meniscal cases were the main stress point. Doubao-Seed-2.0 fell from 2.80 on isolated meniscal injuries to 2.25 on combined injuries, and Qwen3.8-Max fell from 2.80 to 2.45. The two DeepSeek conditions and GLM-5.2 remained comparatively flat, with category means between 2.75 and 2.90. Understandability was the least variable dimension overall (CV 9.8%-12.3%), and GLM-5.2 was most consistent for that dimension (CV 9.8%). Few-shot prompting lowered DeepSeek's accuracy CV from 14.1% to 12.7%, but this gain coincided with the highest mRGL and lowest semantic similarity.

The other expert dimensions showed the same broad pattern with some exceptions. Coverage CV ranged from 11.1% for zero-shot DeepSeek-V4 to 16.6% for Doubao-Seed-2.0, while understandability CV remained within a narrower 9.8%-12.3% interval. GLM-5.2 was again the most consistent on understandability, matching its conservative rewriting style. Stability and mean performance were therefore not interchangeable: the arm with the best average readability had the least stable content scores, while the most consistent readable and fidelity profile belonged to a model with the lowest mean understandability. No arm ranked first on every average and stability measure.

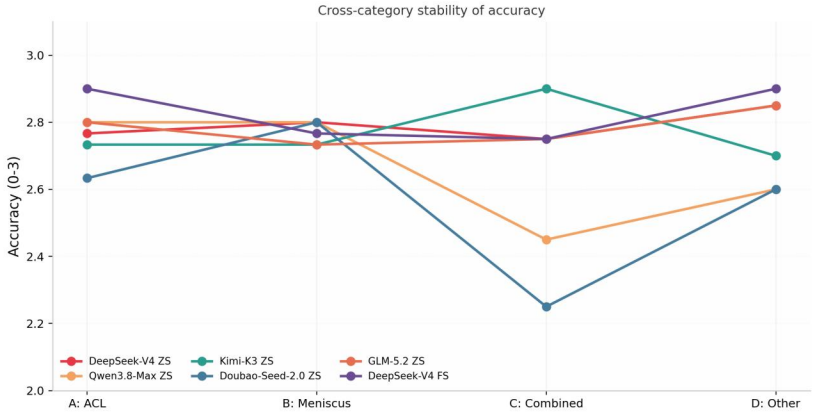

Fig 4. Category-specific accuracy stability across six arms

Note: A flat line indicates similar performance across injury groups; the dips for Doubao-Seed-2.0 and Qwen3.8-Max occur in the combined ACL-meniscal group. ZS, zero-shot; FS, few-shot.

3.5.Inter-rater reliability

Agreement between the two physicians ranged from substantial to almost perfect across all arm-dimension combinations. Weighted kappa was 0.71-0.90 and ICC(3,1) was 0.71-0.91. The highest agreement occurred for understandability in the few-shot DeepSeek-V4 arm (kappa 0.902; ICC 0.907; 95% CI 0.842-0.946). The lowest was also for understandability, but in the Doubao-Seed-2.0 arm (kappa 0.714; ICC 0.714; 95% CI 0.546-0.827). These values indicate that the rubric produced broadly reproducible judgments despite differences in model writing style.

4.Discussion

4.1.Principal findings

This study evaluated 300 summaries generated from 50 guideline-anchored knee MRI cases. The central result was not a statistically distinct accuracy leaderboard: all six arms performed well, and the accuracy differences were not

significant after correction. The practical separation emerged in the dimensions that determine whether an accurate summary is usable. Few-shot DeepSeek-V4 offered the greatest information coverage and the most stable accuracy pattern, but its prose was the hardest to read (mRGL 9.46) and least similar to the source (BERT 0.48). Doubao-Seed-2.0 showed the opposite profile, with the lowest mRGL and highest zero-shot understandability but lower coverage and the largest cross-category fluctuation. GLM-5.2 stayed closest to the radiology wording, although its mean understandability was the lowest.

These differences matter even without a significant accuracy contrast. A one-grade change in mRGL can alter how many readers can understand the text independently, while a 0.24 spread in semantic similarity represents a meaningful difference in how much the model reorganizes the report. The absence of statistical significance should not be interpreted as proof that the arms are identical: 50 paired cases were suited to moderate-to-large effects, not to ruling out small but repeated gaps. For example, few-shot DeepSeek-V4 exceeded Doubao-Seed-2.0 by 0.23 accuracy points, and Doubao-Seed-2.0 produced partially consistent summaries twice as often (32% vs 14%). At large deployment scale, such a difference could still be important.

4.2.Relationship to prior work

The results extend earlier report-simplification studies to a guideline-controlled orthopaedic task and to Chinese flagship systems[9,10,12]. Two observations are consistent with that literature. First, expert judgments and semantic similarity are not interchangeable: few-shot DeepSeek-V4 achieved the highest expert scores while moving furthest from the source wording. Second, no model led on every metric, which argues against evaluating patient-oriented medical rewriting with a single leaderboard number. The high accuracy range observed here (2.60-2.83 out of 3) is compatible with reports that current chatbots seldom invent major facts when rewriting structured medical text[8-10]. The wider dispersion in coverage and readability shows, however, that correctness alone does not describe the patient experience.

The pattern of strong factual preservation alongside uneven accessibility also resembles the exploratory radiology findings of Jeblick et al. and the prompt-learning results of Lyu et al.[9-10]. Kuckelman and colleagues likewise showed that ChatGPT-4 could convert musculoskeletal imaging language into a patient-oriented summary[12]. Our data suggest that Chinese flagship models have reached a similar accuracy plateau for this orthopaedic task. The additional value of the present analysis is to separate coverage from surface readability: GLM-5.2 and Kimi-K3 had comparable expert scores, yet differed by more than one reading-grade level and by 0.12 in semantic similarity.

4.3.Balancing accuracy, readability, and fidelity

The three metric families form a useful trade-off space. Doubao-Seed-2.0 was closest to the readability goal (mRGL 7.15) but gave up information coverage. GLM-5.2 preserved the source most closely (BERT 0.72) but was least understandable on average. Few-shot DeepSeek-V4 improved the qualities that physicians rewarded while worsening the two automated measures. This is important because few-shot prompting is often treated as a low-cost, universally beneficial intervention[8]. In this experiment, the example appears to have encouraged a more detailed and technical register. A demonstration that is comprehensive but written above the target reading level can therefore raise expert scores and lower patient accessibility at the same time. Exemplars should be designed as part of the intervention and evaluated, not copied casually from a preferred output.

One explanation is exemplar anchoring. The demonstration supplied both a writing style and an implicit standard for how much information a summary should contain. The model reproduced the expert-authored balance of precision and terminology, which reviewers valued, but the longer sentences and denser vocabulary increased the calculated reading level. The trade-off may consequently depend on the exemplar rather than on few-shot prompting itself. A deliberately simple exemplar could retain the consistency benefit while reducing the readability penalty. This hypothesis should be tested with controlled exemplar sets and multiple reading levels.

4.4.Why stability should be measured

Mean performance alone concealed the most clinically relevant weakness. When isolated meniscal injuries were replaced by combined ACL-meniscal injuries, accuracy fell by 0.55 points for Doubao-Seed-2.0 and 0.35 for Qwen3.8-Max. These cases require the model to track several structures, grades, and associated findings at once, and an omission in exactly this setting could change how a patient understands prognosis or the range of management options[3,13]. DeepSeek-V4 in both prompt conditions and GLM-5.2 maintained much flatter category profiles. We therefore recommend adding CV, worst-category results, and complexity-stratified means to future evaluation checklists. A patient-facing tool will see the full distribution of report complexity, not only the average case.

4.5.Clinical implications

The findings suggest three safeguards for health systems considering LLM-assisted report explanations. First, the preferred model should match the intended use. Doubao-Seed-2.0 may suit readers who need the simplest language, but its lower coverage requires an omission check. GLM-5.2 may be useful when staying close to the radiologist's text is the priority. DeepSeek-V4, especially with the exemplar tested here, offers a stronger balance of expert-rated coverage and stability, but needs a readability constraint. Second, a clinician should review every patient-facing output because even the strongest arm was not perfectly accurate and a small medical error can have real consequences[8]. Third, readability should be stated in the prompt and verified after generation; expert reviewers can favour a detailed answer that remains too difficult for a patient. These conclusions are consistent with the broader movement toward clinically supervised AI-supported workflows, including the radiomics application described by Kang et al.[30].

A practical workflow would separate drafting from release. The model first generates a summary under an explicit reading-level and completeness instruction. An automated gate then checks mRGL and flags cases with unusually low semantic similarity or missing required fields. A clinician compares the draft with the original report before it is shown to the patient. In this arrangement, the model reduces the writing burden, while the clinician retains responsibility for fidelity and contextualization. Before deployment, the system should also be tested across the same range of case complexity used here so that its safety boundary is defined by its weakest foreseeable scenario.

4.6.Future directions

Several next steps are warranted. The experiment should be repeated with Chinese source reports and Chinese outputs, which better represent routine use and may change the ranking of the systems[16-19]. The simulated bank should be enlarged and compared with de-identified clinical reports to test transfer from controlled cases to real reporting styles. Studies with patients should measure comprehension, anxiety, and decision quality in addition to expert ratings and text metrics. Because web models change over time, repeated snapshots are needed to quantify version drift; the stability framework could be extended to treat model version as a further source of variation. Finally, adaptive prompting could respond to complexity, for example by invoking exemplar-guided instructions for combined injuries. This may capture some of the robustness benefit without accepting the full readability cost observed here.

4.7.Limitations

The findings should be interpreted in light of several limitations. First, the 50 reports were simulated rather than sampled from clinical archives. Simulation allowed close control of guideline alignment and avoided privacy issues, but it cannot reproduce all real-world variation in report quality and style. Second, each model was observed at one time point through a consumer web interface with default settings; later model updates may produce different results, and the reported snapshot covers 10-16 August 2026[16-19]. Third, the ratings came from two sports-medicine physicians. Their substantial agreement is reassuring, but a larger panel including radiologists and patient representatives would improve generalizability. Fourth, bert-base-uncased similarity is a surface-oriented semantic measure and cannot tell safe paraphrase from clinically risky deviation[27]. It should therefore be treated as a fidelity signal, not a safety test. Fifth, all prompts, source reports, and outputs were in English, whereas Chinese-language use is likely to be the dominant real-world scenario. Finally, the few-shot comparison was limited to one model and one exemplar pair, so the observed trade-off cannot yet be generalized to other models or example designs.

5.Conclusions

Across 50 simulated sports-injury knee MRI reports, the five Chinese flagship models produced summaries with similarly high average accuracy, but they differed in information coverage, readability, semantic proximity, and performance on complex cases. The single few-shot intervention improved expert-rated coverage and reduced accuracy variability for DeepSeek-V4 while making the text harder to read and less similar to the source. These results support a multidimensional selection process: define the desired readability level, test performance on combined injuries, monitor stability, and require clinician review before patient release. No model can be considered universally optimal on the evidence available here.

References

[1] MUSAH L V, KARLSSON J. Anterior cruciate ligament tear. New England Journal of Medicine, 2019, 380(24): 2341-2348.
[2] KAEDING C C, LEGER-ST-JEAN B, MAGNUSSEN R A. Epidemiology and diagnosis of anterior cruciate ligament injuries. Clinics in Sports Medicine, 2017, 36(1): 1-8.
[3] BROPHY R H, LOWRY K J. Management of anterior cruciate ligament injuries: AAOS clinical practice guideline summary. Journal of the American Academy of Orthopaedic Surgeons, 2023.
[4] American College of Radiology. ACR Appropriateness Criteria: acute trauma to the knee[Z/OL]. https://acsearch.acr.org/docs/69436/Narrative/. Accessed August 19, 2026.
[5] NIELSEN-BOHLMAN L, PANZER A M, KINDIG D A, eds. Health Literacy: A Prescription to End Confusion. Washington, D.C.: National Academies Press, 2004.
[6] PAASCHE-ORLOW M K, WOLF M S. The causal pathways linking health literacy to health outcomes. American Journal of Health Behavior, 2007, 31(suppl 1): S19-S26.
[7] Office of the National Coordinator for Health Information Technology. 21st Century Cures Act: interoperability, information blocking, and the ONC Health IT Certification Program. Federal Register, 2020, 85(85): 25642-25961.
[8] SALLAM M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. Healthcare, 2023, 11(6): 887.
[9] JEBLICK K, SCHACHTNER B, DEXL J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. European Radiology, 2024, 34(5): 2817-2825.
[10] LYU Q, TAN J, ZAPADKA M E, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Visual Computing for Industry, Biomedicine, and Art, 2023, 6(1): 9.
[11] AYERS J W, POLIAK A, DREDZE M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Internal Medicine, 2023, 183(6): 589-596.
[12] KUCKELMAN I J, WETLEY K, YI P H, et al. Translating musculoskeletal radiology reports into patient-friendly summaries using ChatGPT-4. Skeletal Radiology, 2024.
[13] KOPF S, BEAUFILS P, HIRSCHMANN M T, et al. Management of traumatic meniscus tears: the 2019 ESSKA meniscus consensus. Knee Surgery, Sports Traumatology, Arthroscopy, 2020, 28(4): 1177-1194.
[14] American Academy of Orthopaedic Surgeons. Management of Anterior Cruciate Ligament Injuries: Evidence-Based Clinical Practice Guideline. Rosemont, IL: American Academy of Orthopaedic Surgeons, 2022.
[15] BEAUFILS P, BECKER R, KOPF S, et al. Surgical management of degenerative meniscus lesions: the 2016 ESSKA meniscus consensus. Knee Surgery, Sports Traumatology, Arthroscopy, 2017, 25(2): 335-346.
[16] DeepSeek-AI. DeepSeek-V3 technical report[Z/OL]. arXiv:2412.19437, 2024.

[17] GLM T, ZENG A, XU B, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools[Z/OL]. arXiv:2406.12793, 2024.
[18] YANG A, YANG B, ZHANG B, et al. Qwen2.5 technical report[Z/OL]. arXiv:2412.15115, 2024.
[19] Team Kimi. Kimi k1.5: scaling reinforcement learning with LLMs[Z/OL]. arXiv:2501.12599, 2025.
[20] MEREDITH S J, RAUER T, CHMIELEWSKI T L, et al. Return to sport after anterior cruciate ligament injury: Panther Symposium ACL Injury Return to Sport Consensus Group. Orthopaedic Journal of Sports Medicine, 2020, 8(6): 2325967120930829.
[21] LANDIS J R, KOCH G G. The measurement of observer agreement for categorical data. Biometrics, 1977, 33(1): 159-174.
[22] KOO T K, LI M Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 2016, 15(2): 155-163.
[23] GUNNING R. The Technique of Clear Writing. New York: McGraw-Hill, 1952.
[24] KINCAID J P, FISHBURNE R P, ROGERS R L, et al. Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Millington, TN: Naval Technical Training Command, 1975.
[25] SENTER R J, SMITH E A. Automated readability index. Dayton, OH: Aerospace Medical Research Laboratories, 1967.
[26] COLEMAN M, LIAU T L. A computer readability formula designed for machine scoring. Journal of Applied Psychology, 1975, 60(2): 283-284.
[27] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2019: 4171-4186.
[28] WILCOXON F. Individual comparisons by ranking methods. Biometrics Bulletin, 1945, 1(6): 80-83.
[29] BENJAMINI Y, HOCHBERG Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 1995, 57(1): 289-300.
[30] KANG H, LI X, FENG M, et al. Applying Radiomics and Deep Learning to Investigations into Seafarers' Health Status. Applied Artificial Intelligence Research, 2025, 1(3).