

Hybrid NLP + ChatGPT System for Automated Grading of Short-Answer Radiology X-ray Quizzes

Dr. Tan Xu Hao Isaac, Dr Thida Win, Prof Tang Phua Hwee

Aims/Background

Short-answer X-ray quizzes are crucial for radiology training, yet manual marking is time-consuming and inconsistent. While NLP/LLM advances aid other assessments [1, 2], radiology quizzes' unique lexical and laterality nuances remain unaddressed.

We designed a Python pipeline utilizing fuzzy string matching, a ChatGPT-assisted synonym dictionary, and rule-based checks for automated marking.

Diagnostic accuracy and agreement with expert human grading was evaluated by quantifying concordance. Secondary aims explored how answer length, laterality, and terminology drive disagreement, guiding future refinement.

Methods & Materials

We analyzed 840 anonymized responses from 28 residents across 30 FRCR-style X-ray quizzes, independently scored by an expert. Our automated pipeline standardized answers, then processed them sequentially using three Python modules:

- Fuzzy matching (fuzz.WRatio for string similarity) [3]
- Synonym dictionary matching (ChatGPT-assisted dictionary from 40 sets of pre-existing model answers),
- Rule-based penalties (laterality/anatomical conflicts).

System performance was evaluated across similarity ratios; 65% was chosen as the optimal threshold. Agreement (Cohen's κ), accuracy (diagnostic metrics), directional bias (McNemar's), and word-count prediction of disagreement (logistic regression) were quantified.

Discussion/Results

Our system achieved 97.5% overall concordance with the human grader (819/840 answers, $\kappa=0.892$, substantial agreement), demonstrated high sensitivity, specificity, and predictive values (Table 1).

Metric	Value
Total Concordance (%)	97.5%
Cohen's Kappa (Overall)	0.892
Kappa 95% CI (0.680, 1.000)	
Sensitivity	0.984
Specificity	0.919
Positive Predictive Value (PPV)	0.988
Negative Predictive Value (NPV)	0.895
False Positive Count (FP)	9
False Negative Count (FN)	12
McNemar's Test	p = 0.664

Table 1: Overall Performance Summary

Agreement was perfect ($\kappa=1.00$) for single-word answers and remained high ($\kappa=0.927$) for two-word responses but declined with increasing word count (Table 2). Logistic regression confirmed each additional word increased mismatch odds by 47.5% (OR 1.475, $p < 0.001$).

Word Count	Responses	Cohen's Kappa	Agreement
1 word	429	1.00	100%
2 words	75	0.927	97.33%
3 words	122	0.741	94.26%
4 words	138	0.807	94.23%
5 words	47	0.632	93.62%
6 words	14	0.759	92.86%
7 words	8	N/A*	87.50%
> 7 words	7	1.00	100%

Table 2: Agreement Metrics Stratified by Answer Length

We identified 65% fuzzy-matching threshold as optimal, to maintain high agreement while being the strictest not statistically inferior to peak performance (e.g outperforming 70%, $p = 0.0013$) (Figure 1).

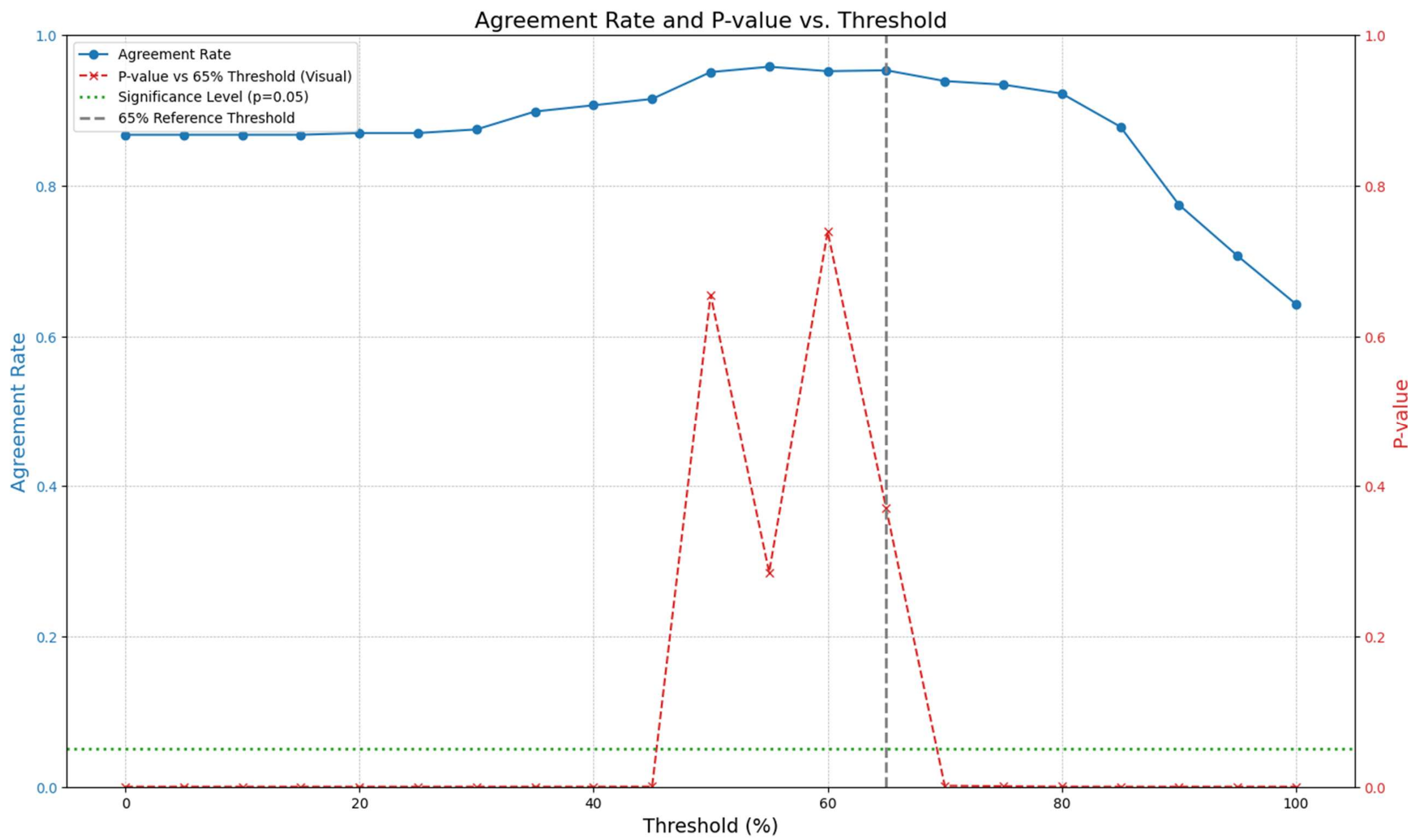


Figure 1. Automated–human agreement across fuzz.WRatio similarity thresholds. Agreement peaks between 50% and 65% for all 840 answers; 65% was adopted as the operational threshold because it lies on a performance plateau confirmed by statistical testing.

Most discrepancies (71.4%) clustered in the top five problematic questions. False positives often involved partially correct/generic answers. False negatives stemmed from descriptive variation, granularity issues, multi-finding answers, or complex lexically-distant correct answers.

Our hybrid pipeline rivals expert human performance, surpasses many LLM-only grading tools, and maintains transparency [4]. This empirically optimized approach provides a reproducible, stringent benchmark [5]. While accuracy declines with longer answers, highlighting lexical limitations, integrating concept-level embeddings (e.g., SapBERT [6]) and partial-credit logic should mitigate this. Discrepancy analysis also offers clear targets for dictionary expansion and rule refinement.

Limitations & Future Directions

Limitations include a single human grader (precluding inter-rater reliability), potential synonym omissions, non-deterministic ChatGPT suggestions, binary scoring (ignoring partial credit), and limited generalizability beyond X-ray quizzes to other modalities or assessment types.

Future enhancements aim for: ontology-powered named-entity recognition (RadLex), transformer encoding for semantic comparison, partial-credit rubrics, annotated-image feedback, a user-friendly GUI for management, and multi-institutional validation across imaging modalities.

Conclusion

This novel study demonstrates the high performance and feasibility of a transparent hybrid NLP and ChatGPT framework for consistently and rapidly grading short-answer X-ray quizzes. Achieving $\geq 97\%$ agreement with human grading at a 65% similarity threshold (perfect for one-word, near-perfect for two-word answers), this approach shows strong potential as a radiology education assessment tool, warranting broader implementation studies.

References

- Burrows S, Gurevych I, Stein B. The Eras and Trends of Automatic Short Answer Grading. Int J Artif Intell Educ. 2015 Mar 1;25(1):60–117.
- Masters K. Artificial intelligence in medical education. Med Teach. 2019 Sep;41(9):976–80.
- Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. Soviet Union; 1966. p. 707–10.
- Grévisse C. LLM-based automatic short answer grading in undergraduate medical education. BMC Med Educ. 2024 Sep 27;24(1):1060.
- Vanallemeersch T, Vandeghinste V. Assessing linguistically aware fuzzy matching in translation memories. In: Proceedings of the 18th Annual Conference of the European Association for Machine Translation. 2015.
- Abdulnazar A, Roller R, Schulz S, Kreuzthaler M. Unsupervised SapBERT-based bi-encoders for medical concept annotation of clinical narratives with SNOMED CT. Digit Health. 2024 Oct 21;10:20552076241288681.