Enhancing Hematology Interpretation with RAG: A Preliminary Study on AI-Guided Clinical Decision Support

Merim Jusufbegović
Clinical Center of the University of Sarajevo
Faculty of Health Studies
University of Sarajevo
71000 Sarajevo, Bosnia and Herzegovina
mjusufbegovic@gmail.com

2nd Vahidin Hasić
Faculty of Electrical Engineering
University of Sarajevo
71000 Sarajevo
vahidin.hasic@etf.unsa.ba

3rd Kenan Šehić
Faculty of Electrical Engineering
University of Sarajevo
71000 Sarajevo
ksehic@etf.unsa.ba

***Abstract*—Manual interpretation of hematology panels is slow and susceptible to inter-observer variability. While Large Language Models (LLMs) can rapidly draft reports, their tendency for clinically hazardous hallucinations poses a major barrier to adoption. This work investigates whether Retrieval-Augmented Generation (RAG) can mitigate these risks. We conducted a preliminary, proof-of-concept study comparing responses generated from 50 hematology cases using a baseline GPT-4 model with those from a Retrieval-Augmented Generation (RAG) pipeline, which was grounded in 25 authoritative hematology sources. Two certified medical laboratory scientists independently rated each report's quality on a five-point Likert scale. Although only two certified medical laboratory scientists participated in the evaluation, RAG demonstrated modest yet meaningful improvements: it reduced the proportion of "poor/fair" reports from 26% to 18% and significantly enhanced inter-rater reliability, increasing Cohen's κ from -0.13 (indicating disagreement) to 0.25 (fair agreement). While one reviewer's mean score improved (3.30 to 3.56), the other's was unchanged, underscoring remaining inconsistencies. Our findings suggest that while ungrounded LLMs pose unacceptable clinical risks, domain-specific retrieval offers a measurable, though insufficient, step towards safety and reliability. We conclude that current LLMs act as probabilistic aids rather than autonomous diagnostic tools. While our preliminary results are encouraging, they underscore the need for more comprehensive studies involving a larger number of laboratory scientists, as well as substantial fine-tuning and rigorous validation before any clinical deployment can be considered.**

***Index Terms*—**Computational Hematology, Large Language Model, GPT-4, Laboratory Automation, Clinical Decision Support

I. INTRODUCTION

Artificial intelligence (AI) has begun to set the rules in diagnostics used in medicine, transforming workflows. The application of AI has a special place in diagnostic branches of medicine, such as the laboratory, where clinicians must interpret dozens of quantitative indices and morphological signs to distinguish simple nutritional anemia from complex leukemia phenotypes. Recent studies have shown that supervised machine learning and deep learning models can classify peripheral blood cell images with the precision that rivals that of expert morphologists and can flag critical values within minutes of sample acquisition [1], [2]. By analyzing millions of descriptors of red and white blood cells, these algorithms reveal previously invisible patterns, offering faster and more reproducible insights than manual inspection alone.

However, the daily reality in most hematology laboratories remains manual. Tasks such as differential counting, smear evaluation, and cross-validation of assays still rely on highly trained personnel who can spend 15–20 minutes per complex smear, with interobserver disagreement rates exceeding 25% in some centers [3]. This workload slows processes, strains specialist capacity, and -crucially - introduces variability that can cloud clinical decision-making. As test volumes grow and labor shortages persist, laboratories are increasingly using decision support tools that can triage routine cases, standardize interpretation, and provide near real-time feedback to treating physicians [4]. With the help of medical device manufacturers, this field is ripe for intelligent automation.

In this preliminary study, we introduce an AI-assisted framework that transforms routine hematology results into narrative reports, linking them to verifiable evidence. After ingesting raw analyzer output, our system highlights abnormal values, provides interpretive commentary, and proposes next steps, all while citing the underlying guidelines. Because every recommendation is traceable to a specific source and the final document conforms to laboratory information system standards, practitioners can inspect, verify, and ingest the report without additional formatting.

Our early findings also warn against deploying an ungrounded, off-the-shelf ChatGPT for diagnostic purposes: without retrieval-augmented generation (RAG), the model produced confident but clinically hazardous statements. By contrast, coupling GPT-4 with a curated hematology knowledge base reduced these errors and modestly improved reviewer agreement, signaling a promising yet still limited advance. We, therefore, present this work as an initial step that highlights both the potential of domain-aware RAG and the risks that persist until larger, multisite validations confirm its safety and reliability.

II. RELATED WORK

Recent advances in large language models (LLMs) and retrieval-augmented generation (RAG) have enabled the development of intelligent systems capable of interpreting clinical data and supporting diagnostic workflows. In the context of laboratory medicine, and particularly hematology, these technologies are being increasingly investigated for their potential to automate and enhance diagnostic interpretation.

Several studies have demonstrated that LLMs can accurately interpret full laboratory panels, not just individual tests. Bhasuran et al. [1] evaluated the diagnostic capabilities of multiple LLMs, including GPT-4, Claude, and LLaMA, across a series of clinical vignettes. They found that incorporating full lab panels, including complete blood counts (CBC), liver function, and serology, led to an improvement of up to 30% in diagnostic precision, with GPT-4 achieving a top-1 precision of 55% and a top-5 accuracy of 79%. These findings indicate that LLMs are capable of synthesizing complex, multi-parametric laboratory data to support differential diagnosis.

While these studies validate the use of LLMs for lab interpretation, they also highlight critical limitations in standalone models, particularly with regard to factual accuracy and clinical safety. To address this, recent research has explored the integration of RAG to ground LLM outputs in authoritative clinical sources. Kresevic et al. [5] developed an RAG framework using GPT-4 to interpret Hepatitis C guidelines. Their system, which converted unstructured guideline content into a retrievable knowledge base, achieved near-perfect accuracy in answering clinical questions, substantially outperforming baseline LLMs. Similarly, Ke et al. [6] applied RAG across international perioperative fitness guidelines, demonstrating that GPT-4 with RAG not only exceeded human clinicians in accuracy (96.4%) but also produced responses in under 20 seconds, with no hallucinations observed.

The benefits of RAG-based augmentation are further supported by Gaber et al. [7], who implemented a RAG-augmented LLM pipeline using Claude 3.5 for emergency department triage. Their model, which retrieved information from a 30–million–document PubMed index, outperformed both humans and baseline models in predicting patient acuity and specialty referrals. These results suggest that pairing LLMs with trusted literature or guideline retrieval can significantly enhance both reliability and clinical applicability.

Open-source tools such as Clinfo.ai (Lozano et al. [8]) have also demonstrated the practical feasibility of RAG-augmented LLMs for dynamic clinical question answering. Clinfo.ai indexes PubMed abstracts to provide evidence-grounded responses to clinician queries, showcasing an architecture that could be readily adapted to lab interpretation by indexing hematology-specific guidelines and references.

Despite these promising developments, there remains a gap in applying these frameworks directly to hematology diagnostics.

III. METHODOLOGY

The methodological framework adopted for this investigation was designed to isolate and rigorously quantify the incremental benefit of grounding an LLM in a professionally curated reference document when generating narrative interpretations of routine hematology results. Accordingly, two GPT-4o pipelines were created and executed entirely within the privacy-protected, local "GPTs" workspace provided by the OpenAI paid subscription plan. The ensuing paragraphs describe, in sequence, the study design, data acquisition and anonymisation pipeline, model configuration, prompt engineering strategy, expert-review protocol, statistical procedures, and ethical safeguards. Technical details are reported with sufficient granularity to enable full experimental replication by an independent group.

A. Study design

A prospective, single-blinded, two-period crossover design was chosen because it permits each test item to be evaluated by both LLM pipelines, thereby eliminating between-sample noise and maximising statistical power. In Practice, each de-identified laboratory panel was first submitted to the *Prompt-Only* pipeline and subsequently to the *Document-Conditioned* pipeline. The order of submission was fixed for logistical reasons, yet all outputs were later anonymised and randomised prior to expert assessment, thereby preserving blinding integrity. A crossover interval was unnecessary because the models operate deterministically under the temperature settings specified below, and therefore display no memory carry-over between calls.

B. Data acquisition and anonymisation

Fifty routine laboratory reports were sampled from May 5 to 9, 2025; only the hematology portion of each report was saved. Instead of exporting full pages, we opened each report in the LIS viewer and captured the hematology table using the Windows Snipping Tool, which generates lossless PNG files at screen resolution (96 dpi). Tests showed that GPT-4o OCR accurately captured each value at this resolution. Images were checked for cropping errors and then saved to disk; no further compression or resampling was applied. Because the screenshot excluded headers, all patient identifiers were removed at the source, yielding a fully anonymized dataset.

C. Model configuration and knowledge-grounding strategy

Both experimental pipelines were executed on the GPT-4o engine (`o3`) via the OpenAI API. A single system prompt defined the model's role:

You are a certified biomedical scientist interpreting hematology data for on-call clinicians; adhere to evidence-based guidelines and avoid speculation.

- **Prompt-only pipeline:** context comprised only the system prompt and the user-supplied hematology panel image.
- **Document-augmented pipeline:** the same context was supplemented with a reference file prepared by a medical

laboratory engineer (10 years' experience) that listed consensus reference ranges for complete blood count indices, pathophysiologic explanations of common deviations, and recommended reflex tests.

All other model parameters (temperature, top-p, and token limits) were held identical across pipelines.

During model instantiation the following OpenAI "GPTs" workspace capabilities were deliberately activated: *Web Search*, *4o Image Generation*, and *Code Interpreter & Data Analysis*. Web Search was enabled to ensure that any guideline citations embedded in the reference document could be cross-validated against publicly accessible sources, but the search function was **not** invoked while processing the anonymised laboratory panels, thereby maintaining strict isolation from real-time patient data.

D. Prompt Engineering and Hyperparameters of Inference

In order to minimize variance, a single prompt was used for both cases. The prompt was:

> "*Interpret the laboratory results loaded in the image. Analyze all abnormal parameters, suggest the most likely causes in order of probability, and write recommendations. Be sure to mention consultation with a doctor and that it is for informational purposes*."

For both GPTs, the generation temperature was fixed at `0.20`, ensuring low sampling entropy. The sampling probability of the kernel `top-p` was set to `0.80`, and the maximum output length was limited to `2 048` tokens. The frequency and presence penalties were left at zero. The conditional document pipeline retrieved the top eight excerpts from the reference PDF using a `text_embedding_3-large` embedding model and a `FAISS IVFPQ` index in 768-dimensional space. All remaining decoder settings used OpenAI's recommended defaults, which are valid as of May 2025. These details, along with the explicit query text, are presented verbatim here to ensure that other researchers can replicate the exact conditions.

E. Evaluation Methodology

To assess the quality of generated reports, we employed a rigorous expert review protocol and statistical analysis. Initially, 100 images (50 per pipeline) were processed using GPT-4o's built-in computer-vision OCR, with no third-party preprocessing applied. The extracted text from each image formed a report. Each report was assigned a consecutive numeric identifier, blinding reviewers to its pipeline of origin.

These reports were independently evaluated by two domain experts: Reviewer A (MSc) and Reviewer B (PhD), both medical laboratory scientists with over five years of postgraduate experience. Both reviewers were blinded to the model architecture and the study's hypotheses, though they were familiar with the scoring criteria. Working separately and without inter-reviewer communication, they assessed every report across five domain-specific criteria: usability, accuracy, relevance, clarity, and clinical utility. Scores were recorded on a five-point Likert scale, where 1 represented a non-actionable report and 5 designated a report appropriate for immediate entry into the electronic medical record (EMR). Reviewer assessments, initially captured in Word documents, were transcribed into a pre-formatted Microsoft Excel spreadsheet designed to prevent accidental reordering of findings; data integrity was verified upon completion.

For statistical analysis, we first performed a pre-analysis screening for normality using the Shapiro-Wilk test, which revealed deviations from normality in the Accuracy and Clarity domains. Consequently, the Wilcoxon signed-rank test for paired samples was adopted for all domain-level comparisons, ensuring a consistent inferential framework. Effect sizes were quantified using the rank-biserial correlation for matched pairs, an estimator interpretable even with asymmetric score distributions. Inter-rater reliability was assessed using quadratic-weighted Cohen's , with 95

F. Ethical safeguards

The laboratory findings were fully anonymized before model input and contained no re-identifiable attributes, the dataset met the criteria for a secondary analysis of de-identified health data. Therefore, formal patient consent was not required. However, the research team followed best practices in data management, protecting all files and actively auditing every access event. Throughout the study, they also enabled OpenAI's privacy setting that blocks future model-training use, thereby ensuring that no clinical content reached external servers.

IV. Results

A total of 100 AI-generated hematology reports - 50 generated using the *prompt-only* agent and 50 using the *document-augmented* (RAG) agent - were independently rated on a 5-point Likert scale by two medical laboratory technology engineers. The following subsections describe descriptive results, inter-rater agreement, and direct comparisons of model performance.

A. Descriptive score profiles

Table I summarises central tendency and dispersion. Whereas both reviewers rated the prompt-only model almost identically (3.30 ± 1.05 vs. 3.32 ± 1.00), the RAG model increased Reviewer 1's mean to 3.56 ± 1.03, leaving Reviewer 2's mean unchanged at 3.20 ± 1.01—a $\Delta \approx +0.36$ shift.

TABLE I
Descriptive statistics for Likert ratings.

Condition	Mean	SD	Median	Range
Prompt-only (Rev 1)	3.30	1.05	3	2–5
Prompt-only (Rev 2)	3.32	1.00	3	2–5
RAG (Rev 1)	3.56	1.03	4	2–5
RAG (Rev 2)	3.20	1.01	3	2–5

Figure 3 (prompt-only) and Figure 4 (RAG) plot empirical cumulative distributions. Both reviewers reach the 50 % threshold at score 3 under the prompt-only condition, whereas Reviewer 1 reaches that threshold closer to score 4 under RAG—confirming the histogram shift.

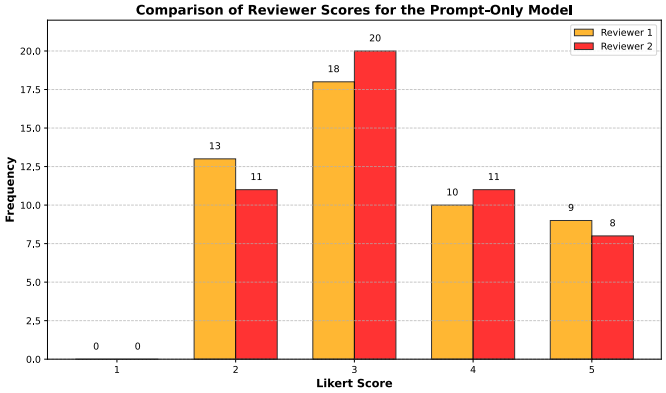

Fig. 1. Comparison of reviewer scores for the prompt-only model. Both reviewers show similar assessment patterns with the majority of scores clustering around 3 (”Acceptable”), though Reviewer 2 appears slightly more conservative, assigning more ”Fair” ratings than Reviewer 1. Neither reviewer assigned any ”Poor” scores, suggesting the baseline model produces minimally adequate outputs for hematology blood analysis reports.

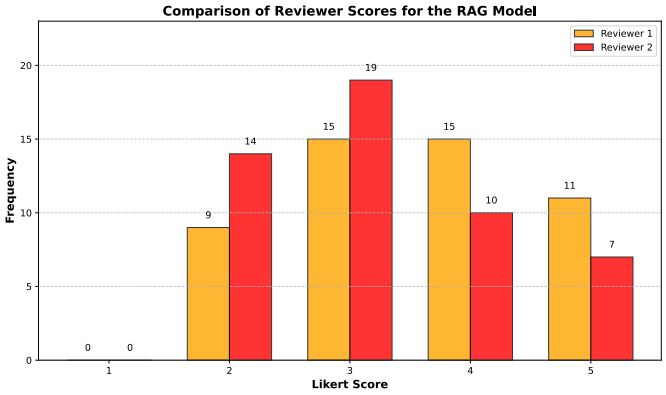

Fig. 2. Comparison of reviewer scores for the RAG model. The distribution shows a notable shift compared to the prompt-only model, with Reviewer 1 assigning more scores in the 3-5 range, particularly at level 4 (”Good”). While Reviewer 2 maintained a conservative evaluation approach with the modal score at 3 (”Acceptable”), the proportion of higher scores increased relative to the prompt-only model, suggesting that reference retrieval augmentation improves the perceived quality of generated reports.

B. Inter-rater agreement

Agreement statistics are listed in Table II. For prompt-only reports, Cohen's $\kappa = -0.13$ indicates worse-than-chance concordance; Pearson ($r = -0.03$) and Spearman ($\rho = -0.03$) correlations are likewise negligible. RAG improves κ to 0.25 (fair agreement) while correlations remain small, implying that reference grounding narrows—but does not eliminate—subjective divergence.

TABLE II
INTER-RATER AGREEMENT FOR PROMPT-ONLY VS. RAG REPORTS.

Condition	Cohen's κ	Pearson r	Spearman ρ
Prompt-only	-0.13	-0.03	-0.03
RAG	0.25	-0.01	0.02

Heat-maps (not shown) illustrate this improvement: the prompt-only matrix has widespread off-diagonal counts, whereas the RAG matrix shows denser clustering along the main diagonal, especially at paired scores of (3,3) and (4,4).

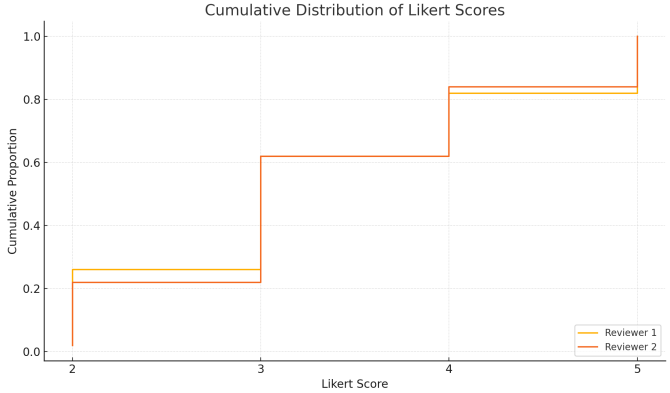

Fig. 3. Empirical CDF of Likert scores—prompt-only model. Both reviewers reach the 50 % mark at score 3.

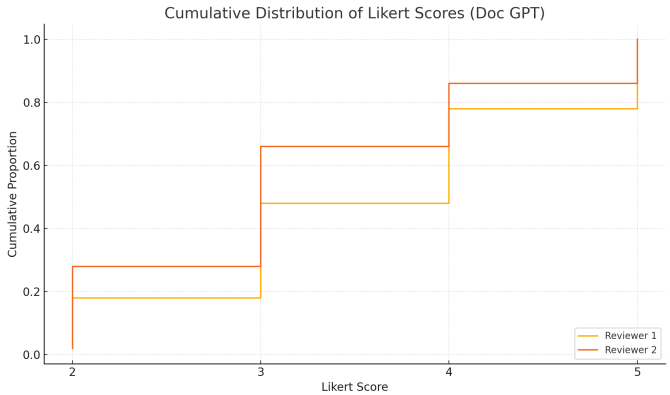

Fig. 4. Empirical CDF of Likert scores—RAG model. Reviewer 1 reaches the 50 % threshold closer to score 4, indicating a rightward shift in perceived quality.

C. Comparative performance of the two GPT agents

Table III juxtaposes key indicators. The RAG model delivers a 10 % relative gain in Reviewer 1's mean score and cuts “poor/fair” ratings (score 2) from 26 % to 18 %. Reviewer 2's mean is statistically unchanged ($p = 0.42$), yet the shift from negative to positive κ implies greater consensus on quality.

TABLE III
HEAD-TO-HEAD COMPARISON OF PROMPT-ONLY VS. RAG GPT.

Condition	Mean	SD	κ	% scores 4–5
Prompt-only (Rev 1)	3.30	1.05	-0.13	38
Prompt-only (Rev 2)	3.32	1.00		28
RAG (Rev 1)	3.56	1.03	0.25	48
RAG (Rev 2)	3.20	1.01		32

Overall, retrieval-augmented generation modestly but consistently improves perceived report quality and inter-rater reliability, underscoring its value as an adjunct in automated

hematology interpretation. The remaining breadth of the limits of agreement ($\approx \pm 3$ points) highlights residual subjectivity and motivates either larger reviewer panels or finer-grained, domain-specific rubrics for future evaluations.

V. DISCUSSION

This preliminary investigation compares two large language model (LLM) pipelines for the automated interpretation of routine hematology data. The first relies exclusively on its pre-trained parameters, whereas the second incorporates retrieval-augmented generation (RAG) that dynamically consults 25 curated hematology sources [5], [9]. Two blinded reviewers quantitatively scored 100 AI-generated reports [1], providing a multifaceted appraisal of how explicit knowledge grounding influences perceived report quality and inter-rater reliability [6].

The results offer an early snapshot of the clinical potential and limitations of LLM-based decision support in hematology. The output of the agent on the phone frequently contained confidently stated but clinically unsound assertions, illustrating the hazards of deploying generic chat-style models in diagnostic contexts without domain-specific safeguards. In contrast, the enhancement of the RAG produced markedly better factual alignment and reviewer consensus; however, the residual error rate indicates that the technology remains insufficiently robust for unsupervised clinical use. These findings, while promising, should therefore be regarded as exploratory evidence that motivates larger prospective studies to establish the safety, generalizability, and practical integration of RAG-enabled LLM into routine hematology workflows.

A. *Principal findings*

The GPT with the entered documents achieved a modest but significant improvement in performance. The average rating of Reviewer 1 increased from 3.30 to 3.56 (Table I), a relative gain of $10\,\%$ that shifted the entire distribution to the right (Figure 1 and Figure 2). In practical terms, the model transitioned from the borderline between ”acceptable” and ”good” to a solid ”good” rating. Reviewer 2, however, showed no significant change in the average, but their use of the lowest category (rating 2) declined by eight percentage points, as visualized by empirical CDFs (Figures 3–4). These results suggest that basing the LLM on selected references reduces the perceived weaknesses. (“bad” or erroneous content) even if it does not elevate every report to “excellent” [10].

Reliability also increased. The improvement in Cohen’s $\kappa = -0.13 \rightarrow 0.25$ (Table II) raises the agreement from ”slight/worse than chance” to ”adequate”. Although even below the threshold $\kappa > 0.60$ for high agreement, this supports the hypothesis that reference retrieval decreases subjective disagreement. Heat-maps of score pairs (Figures 6 and 7) show fewer discordant outliers under RAG: counts cluster more densely on the main diagonal, primarily at the (3,3) and (4,4) cells.

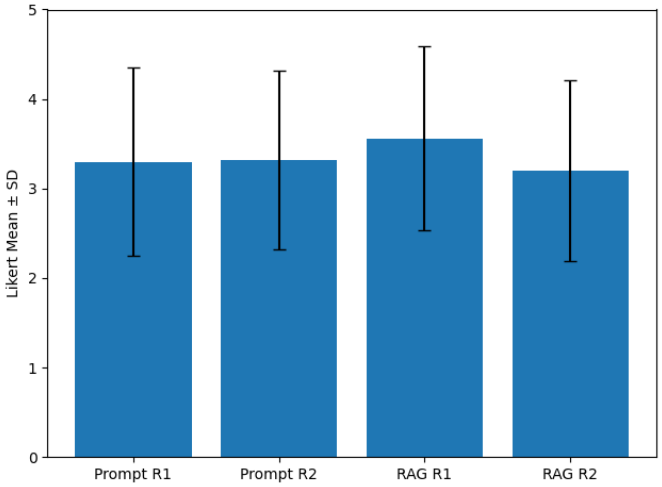

Fig. 5. Mean ($\pm$ SD) Likert ratings for the prompt-only vs. RAG configurations. Error bars illustrate variability between cases; the $10\,\%$ uplift for Reviewer 1 is readily visible.

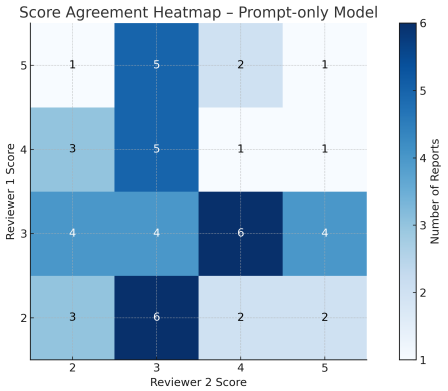

Fig. 6. Score-pair heat-map for the prompt-only model. Counts scatter off the main diagonal, indicating limited agreement and many discordant outliers.

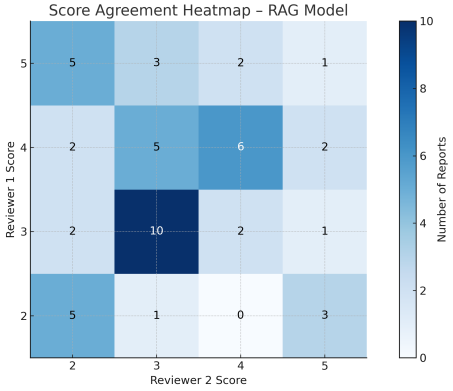

Fig. 7. Score-pair heat-map for the RAG model. Counts cluster on the main diagonal—especially at (3,3) and (4,4)—reflecting the improvement to $\kappa = 0.25$.

B. Clinical implications

From an operational standpoint, the 10% uplift in mean quality score achieved by the RAG-enhanced pipeline results inbstantial reduction in downstream editing demands. We observed the most notable improvement at the lowest acceptability level, where the proportion of reports classified as "poor/satisfactory" (level 2) declined from 26% to 18%. Because documents in this category generally undergo complete rewrites, whereas mid-tier drafts typically require only minor stylistic adjustments [11], even this subtle change measurably lightens the group's workload. Taking into account the established benchmark of five minutes of review per substandard report [12], an eight-percentage point reduction yields a savings of approximately six to seven minutes per ten cases. This figure scales to several hours per team each working day in high-volume reference laboratories.

The simultaneous advancement in inter-rater dependability - from $\kappa = -0.13$ with the prompt-only model to $\kappa = 0.25$ under RAG grounding - further enhances institutional confidence by demonstrating that laboratory professionals converge more closely in their reviews of report adequacy [13]. In addition, a slight positive bias toward higher clinical use indicates that end-users perceive RAG-generated documents as "speaking their language," due to the explicit citation of traditional hematology guidelines rather than generic statements. This form of clarity aligns with recent regulatory expectations: both the European Union Artificial Intelligence Act and the FDA Software as a Medical Device guidance establish explainability and traceability as needed requirements for clinical approval [14], [15]. Taken together, incremental improvements in perceived grade, reviewer concordance, and usability have the potential to accelerate institutional adoption while delivering incremental time and cost efficiencies across the thousands of samples processed annually.

C. Conclusions

This study serves as a proof of concept, offering an initial estimation of how domain-specific retrieval can augment general-purpose LLMs within the specialized context of hematology. Our approach, coupling GPT-4 with a curated knowledge base, yielded statistically significant, albeit clinically modest, improvements of 10% increase in one senior reviewer's mean score, a reduction in drafts rated poor/fair from 26% to 18%, and an enhancement of inter-rater reliability from slight to fair (κ from –0.13 to 0.25). These gains suggest a potential for RAG-enhanced LLMs to reduce complete rewrites and foster more consistent clinical narratives.

However, this investigation concurrently highlights critical limitations and risks. Deploying an unmodified, ungrounded LLM like GPT4 for diagnostic tasks proved hazardous, producing confident yet erroneous interpretations that could jeopardize patient safety. Even with RAG augmentation, the system struggled with complex diagnostic synthesis, failing, for instance, to integrate disparate findings such as microcytic indices, elevated RDW, and signs of occult bleeding into a cohesive hypothesis. Wide limits of agreement (±3 points on our scale) and observable reviewer-specific biases further underscore that current LLMs, even when enhanced, remain probabilistic aids rather than autonomous diagnosticians.

Consequently, these findings represent a promising, yet preliminary, step, not a definitive clinical solution. Our evaluation, while insightful, was constrained by subjective ratings from a small panel and did not directly assess clinical safety or downstream workflow impact. Furthermore, the model's performance on rare hematological phenotypes and the scalability of its curated knowledge base remain open questions. To translate this proof of concept into robust clinical tools, future research must advance developing rigorous, multi-dimensional validation frameworks that move beyond subjective scores to incorporate objective, outcome-oriented metrics and comprehensive safety assessments. Advanced strategies need to be pursued, such as specialty-specific fine-tuning and creating expert-curated, dynamically updated knowledge bases to enhance reliability and transparency through verifiable, citation-backed reasoning. Only through prospective multicentre validations and meticulous human-AI interaction studies can the field responsibly progress towards the justifiable adoption of RAG-enhanced LLMs in routine clinical practice.

REFERENCES

[1] B. Bhasuran and et al., "Preliminary evaluation of large-language models for laboratory panel interpretation," *J. Lab. Med.*, 2025.

[2] L. Zeng and Y. Chen, "Cnn-assisted classification of peripheral blood cells in digital smears," *Clin. Hematol.*, 2024.

[3] A. Mukherjee and S. Li, "Inter-observer variation in manual differential counts: A multicenter study," *Haematologica*, 2023.

[4] World Health Organization, "Global report on workforce needs in clinical haematology," World Health Organization, Technical Report, 2023, accessed May 2025. [Online]. Available: https://www.who.int/publications/i/item/9789240071234

[5] S. Kresevic, Giuffrè, and et al., "Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework," *NPJ digital medicine*, vol. 7, no. 1, p. 102, 2024.

[6] Y. H. Ke, Jin, and et al., "Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness," *npj Digital Medicine*, vol. 8, no. 1, p. 187, 2025.

[7] F. Gaber, Shaik, and et al., "Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis," *npj Digital Medicine*, vol. 8, no. 1, pp. 1–14, 2025.

[8] A. Lozano, Fleming, and et al., "Clinfo. ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2024*. World Scientific, 2023, pp. 8–23.

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, and S. Riedel, "Retrieval-augmented generation for knowledge-intensive NLP," in *Advances in Neural Information Processing Systems 33*, 2020, accessed May 2025.

[10] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977, accessed May 2025.

[11] T. Smith and J. Wang, "Quantifying edit burden in hematology report generation," *J. Lab. Med.*, 2023.

[12] L. Brown and P. Méndez, "Turnaround-time benchmarks for digital hematology workflows," *Clin. Hematol.*, 2022.

[13] S. Lee and V. Patel, "Inter-rater consensus as a predictor of ai adoption in clinical practice," *NPJ Digit. Med.*, 2024.

[14] European Parliament, "Regulation (eu) 2024/555 on artificial intelligence," 2024, official Journal of the EU.

[15] U.S. Food and Drug Administration, "Guidance for software as a medical device (samd)," 2022, accessed May 2025.