Comparative Performance of Multimodal Large Language Models in Acute Leukemia Morphology

Miranda Lin, M.D.[1], Yoshito Nishimura, M.D.[2], Lauren Pomerantz, M.D. [1], Hadil Zureigat, M.D.[2], Vikram Puram, M.D. [2], Jesse Fajnzylber, M.D. [1], Jordan Nunnelee, M.D. [2], Katherine Talbott, M.D. [3], Richard Godby, M.D. [2]

[1]Department of Internal Medicine, [2]Department of Hematology, [3]Department of Anesthesiology

BACKGROUND

Multimodal large language models (LLMs) can interpret clinical images, but their reliability, calibration, and safety in hematopathology remains uncertain. This study aimed to compare 3 publicly accessible multimodal LLMs for acute leukemia image interpretation using curated peripheral blood (PB) and bone marrow (BM) images.

METHODS

22 deidentified PB/BM images in 5 acute leukemia categories were selected from the ASH Image Bank. ChatGPT 5.2, Gemini 3.0 Pro, and ChatGPT for Clinicians (Models 1, 2, and 3, respectively) evaluated each image using identical zero-shot, non-iterative prompts.

Prompt 1: Smear Evaluation

Pretend that you are a trained hematopathologist who has been asked to review this peripheral blood or bone marrow smear for a human patient. Carefully examine the different features of each cell line, including platelets, white blood cells, and red blood cells. Are there too many or too little of each? Are the cell sizes relative to each other normal? Are there any morphologic abnormalities in any of the cell lines? Any other abnormalities that you can detect? Please report your findings in 5 sentences as follows: first describe the platelets, then describe the white blood cells, then describe the red blood cells, then describe your differential diagnosis based upon your analysis of the peripheral blood smear, and then finally recommend best next steps to care for the patient, including additional labs and treatments.

Prompt 2: Composite Analysis

Please synthesize all previous findings and provide a final diagnostic analysis. Please review all of your previous analyses collectively, provide a summary of the key, recurrent morphologic findings across all smears, state the most likely diagnosis that is common to all provided peripheral smears, and provide a detailed justification for your diagnosis, referencing the specific, cumulative evidence from all the images.

Prompt 3: Reflection

The final diagnosis across the five slides is AML NOS. Please review your interpretations and reflect on why you got the diagnosis correct or wrong.

FIGURE 1. Zero-shot, non-iterative inputted into 3 LLMs.

RESULTS

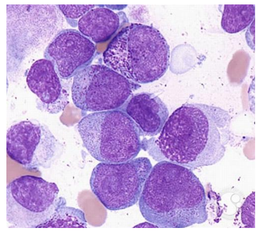

FIGURE 2. Representative image of AML-NOS uploaded to ChatGPT for Clinicians (Model 3) and response.[1]

Platelets are not readily apparent in this field and appear reduced, without conspicuous giant or markedly dysplastic platelet forms.
White blood cells are overwhelmingly abnormal, composed of large blasts with high nuclear-to-cytoplasmic ratios, fine chromatin, irregular to folded nuclei, and moderate pale basophilic cytoplasm containing fine azurophilic granules in some cells; this is strongly supportive of an acute myeloid neoplasm and not a reactive process.

Red blood cells are sparse in this marrow-rich image and are not the dominant abnormality, with no specific red-cell morphologic diagnosis possible from this field alone.

The differential diagnosis is led by acute myeloid leukemia, with morphology raising concern for a myeloblastic process and possibly myelomonocytic/monocytic differentiation, while acute lymphoblastic leukemia is less favored on morphologic grounds; morphology alone cannot reliably establish the exact AML subtype.

Best next steps are urgent hematology management with CBC/differential, CMP, creatinine, uric acid, phosphorus, calcium, LDH, PT/INR, aPTT, fibrinogen, D-dimer, peripheral blood and/or marrow flow cytometry, bone marrow core/aspirate review, cytogenetics/FISH, and molecular testing, plus immediate supportive care for tumor lysis, bleeding/DIC, infection, and leukostasis risk as indicated, with definitive induction strategy based on confirmed subtype.

- Several failure modes identified:
  - Lineage misalignment
  - Anchoring on isolated visual features/artifacts
  - Hallucinations of alternative diagnoses
  - Overconfidence by using more definitive diagnostic language
  - Premature management recommendations
- Correct next management was recommended:
  - Urgent hematology consultation
  - CBC, coagulation panel, CMP, creatinine, uric acid, phosphorus, calcium, LDH, type and screen, peripheral blood flow cytometry
  - BM aspirate/biopsy with cytogenetic and molecular studies
  - Supportive care including tumor lysis prophylaxis and transfusion support as indicated
- All 3 LLMs reflected on the correct diagnosis, identified areas of failure, and provided recommendations for improvement

Category	ChatGPT 5.2 (Model 1)		Gemini 3.0 Pro (Model 2)		ChatGPT for Clinicians (Model 3)	
	Correct/Total	Unifying Diagnosis	Correct/Total	Unifying Diagnosis	Correct/Total	Unifying Diagnosis
AML-NOS	4/5 (80%)	**Acute myeloid leukemia**	1/5 (20%)	No unifying diagnosis	4/5 (80%)	**Acute myeloid leukemia**
AML-MRC/AML-MR	5/5 (100%)	**Acute myeloid leukemia**	4/5 (80%)	**Acute myeloid leukemia**	4/5 (80%)	Acute promyelocytic leukemia
B-ALL NOS	4/4 (100%)	Acute leukemia	3/4 (75%)	Metastatic solid tumor	4/4 (100%)	**Acute lymphoblastic leukemia/B-lymphoblastic leukemia**
B-ALL BCR-ABL like	3/3 (100%)	**B-lymphoblastic leukemia/lymphoma**	3/3 (100%)	Acute promyelocytic leukemia	3/3 (100%)	Acute leukemia
T-ALL	4/5 (80%)	**Acute lymphoblastic leukemia**	1/5 (20%)	Burkitt lymphoma/leukemia	3/5 (60%)	**T-lineage acute lymphoblastic leukemia/lymphoblastic lymphoma**
TOTAL	20/22 (91.0%)		12/22 (54.5%)		18/22 (86.4%)	

TABLE 1. Percent correct in each acute leukemia category. A correct diagnosis was considered "acute leukemia." Each LLM produced a unifying diagnosis after analyzing every image in each category. Correct lineage identified by the LLM is in bold font.

CONCLUSIONS

- Publicly available multimodal LLMs have variable diagnostic accuracy for acute leukemias
- Safety-relevant failure modes present
- Diagnostic accuracy of LLMs are not superior to hematopathologist interpretation
- Diagnostic concordance may be improved with training on disease-specific datasets
  - Ie. plug-ins
- Clinical integration will require further validation studies and expert oversight
- LLMs may be utilized as educational tools and preliminary triage aids
  - Potential in limited hematopathology access settings
- Limitations:
  - Small image set
  - Possible prompt sensitivity
  - Unclear external generalizability to real-world scenarios

REFERENCES

1. Peter Maslak. Acute Myeloid Leukemia with Maturation. 2003; 00002212. ASH Image Bank. © the American Society of Hematology.
2. ASH Image Bank. https://imagebank.hematology.org/
3. Apornvirat S et al. *Ann Diagn Pathol*. Dec 2024;73:152359.
4. Cai X et al. *Digit Health*. Jan-Dec 2024;10:20552076241298503.