MELBA Journal Machine Learning for Biomedical Imaging

Multilingual Hematology Visual Question Answering Dataset

Hajra Malik [1], Hafiza Tooba Aftab [2], Abdul Rehman [1], Mohsen Ali [1], Waqas Sultani [1]

1 Information Technology University, Lahore
2 King Edward Medical University, Lahore

Abstract

Vision Language Models (VLMs) have shown promising capabilities in medical image analysis by jointly understanding visual and textual information for tasks such as Visual Question Answering (VQA). However, existing hematology vision-language resources remain predominantly English-centric, limiting their applicability in multilingual healthcare environments. This challenge is particularly relevant in South Asia, especially in Pakistan, where Urdu is widely spoken, while healthcare information and digital healthcare systems remain largely English-based. To investigate this gap, we surveyed healthcare professionals and identified substantial language mismatches between clinical documentation and patient communication, underscoring the need for multilingual healthcare technologies. To address this limitation, we introduce WBCMor-VQA, a clinically validated bilingual (English–Urdu), morphology-aware VQA benchmark for leukemia and normal white blood cell (WBC) analysis. The benchmark is constructed using morphology-aware annotations from the LeukemiaAttri and WBCAtt datasets and is supported by a domain-specific Urdu hematology dictionary to ensure linguistic consistency and clinical correctness. The final benchmark comprises 110K bilingual question–answer pairs corresponding to 20K leukemic and normal single-cell images. Furthermore, we establish strong baseline results by evaluating multiple open-source VLMs on the proposed benchmark. The proposed resource aims to facilitate the development of accessible and clinically relevant AI systems for multilingual healthcare environments. The dataset is publicly available at: https://doi.org/10.6084/m9.figshare.32727159

Keywords

Leukemia, Hematology, Vision-Language Models (VLMs), Visual Question Answering (VQA), White Blood Cell Morphology, Multi, Urdu Healthcare, Bilingual Dataset

Article informations

https://doi.org/10.59275/j.melba.2026-4182
Volume 3, Received: 2026-06-19, Published 2026-09-21
Corresponding author: waqas.sultani@itu.edu.pk
Special issue: MICCAI Open Data 2026 x MELBA
Guest editors: AT, MS et al.

1. Background

Leukemia, a blood cancer that affects the production and function of blood cells, is commonly diagnosed and assessed by examining White Blood Cells (WBCs). The disease poses a significant global health challenge because it directly impairs the human immune system, compromising its ability to fight infections (Blood Cancer United, 2025). Long-term epidemiological evidence indicates that leukemia remains one of the most prevalent hematologic malignancies in diverse populations, highlighting its continued clinical and public health significance (Badar and Mahmood, 2023). Despite advances in diagnostic techniques and treatment strategies, leukemia continues to contribute substantially to global cancer-related morbidity and mortality. Its aggressive nature is reflected in its disproportionate impact on children and young adults, where it remains among the leading causes of cancer-related deaths in individuals younger than 45 years of age (Campbell and Grondona, 2008).

In Pakistan, hematological malignancies, including leukemia, account for a considerable proportion of the national cancer burden (Tufail and Wu, 2023). Although healthcare services have improved over time, cancer-related mortality remains high due to challenges such as delayed diagnosis, limited public awareness, and restricted access to specialized medical care. Recent estimates report an age-standardized cancer mortality rate of 153.52 deaths per 100,000 individuals (Bhojwani et al., 2026). Among pe-

diatric populations, leukemia is of particular concern, as acute lymphoblastic leukemia (ALL) continues to be one of the most frequently diagnosed childhood cancers in the country (Tufail and Wu, 2023). Beyond the clinical challenges associated with leukemia, access to healthcare information and emerging medical technologies is strongly influenced by language. South Asia, with a population of approximately 2.11 billion people, is characterized by substantial linguistic diversity. Among the major regional languages, Urdu is spoken by nearly 246 million people and serves as an important medium of communication across communities and national boundaries (Poria et al., 2025). As the national language of Pakistan and a widely understood language across South Asia, Urdu plays a critical role in facilitating access to healthcare information and improving communication between healthcare systems and local populations (Campbell and Grondona, 2008).

Despite the widespread use of Urdu, healthcare information, medical records, and digital health technologies in Pakistan continue to rely primarily on English. Differences in literacy and English proficiency across regions and socioeconomic groups create barriers to understanding medical information and interacting effectively with healthcare systems. According to the Education First (EF) English Proficiency Index 2025 rankings, Bangladesh, Pakistan, and India are ranked 62nd, 67th, and 74th globally, suggesting that English-based healthcare and AI technologies may remain inaccessible to a large portion of the population (De Angelis, 2025). To further examine the role of language in clinical communication, we conducted a survey involving healthcare professionals who regularly interact with patients and their families. The results revealed a noticeable gap between the language used in medical documentation and the language commonly used during patient interactions. A majority of physicians (58.6%) reported communicating with patients using a combination of Urdu, English, and local languages, while 34.9% primarily relied on Urdu or other local languages. In contrast, only 6.5% reported communicating primarily in English. Additionally, 75.1% of respondents indicated that patients often or sometimes misunderstand medical explanations, reflecting persistent communication challenges in routine clinical settings. These findings underscore the need for multilingual healthcare resources and language-adaptive medical technologies to improve communication between healthcare providers and patients. A detailed description of the survey methodology, participant demographics, and complete findings is provided in Section 3.1.

Recent advances in Vision-Language Models (VLMs) have enabled artificial intelligence (AI) systems to jointly understand medical images and natural language (Kalpélbé et al., 2025; Hartsock and Rasool, 2024). This capability has accelerated the development of multi-modal healthcare applications, including medical image interpretation, automated report generation, and Visual Question Answering (VQA) systems (Bazi et al., 2023). Among these applications, VQA has gained significant attention because it enables users to interact with medical images through natural-language queries and receive clinically relevant responses (Noshin et al., 2026). Such systems have the potential to improve access to medical knowledge, strengthen clinical decision support, and provide more interpretable AI-assisted healthcare solutions.

Based on these developments, Uni-Hema (Rehman et al., 2026) introduced a comprehensive vision-language framework for digital hematology. The framework utilized morphology annotations from the WBCAtt dataset (Tsutsui et al., 2023), which contains normal white blood cells with detailed morphological attribute labels, to generate morphology oriented VQA samples. In addition, leukemia specific language supervision was incorporated through masked language modeling, allowing joint learning of visual and textual representations. More recently, HemBLIP (van Logtestijn and Manescu, 2026) further demonstrated the effectiveness of VLM approaches for leukemia analysis by generating morphology aware descriptions directly from microscopic blood cell images. Although these studies demonstrate the growing potential of vision-language methods in hematological analysis, existing resources remain predominantly English-centric and are primarily designed for caption generation, representation learning, or automatically generated language supervision. As a result, clinically validated multilingual resources, particularly for Urdu-speaking healthcare settings, remain largely unavailable. Motivated by this limitation, we develop a clinically validated bilingual VQA resource to understand the morphology of normal and leukemic cells.

2. Summary

The proposed **WBCMor_VQA** dataset has been developed using morphology-aware annotations from the LeukemiaAttri and WBCAtt datasets (Rehman et al., 2024; Tsutsui et al., 2023). These datasets provide detailed morphological information for leukemia and normal white blood cell images, allowing the generation of clinically relevant visual question-answer (VQA) pairs. For dataset curation, we first extracted single-cell images from the full field-of-view microscopy images and, following the morphology-oriented VQA generation strategy introduced in Un-Hema (Rehman et al., 2026), constructed English question–answer pairs and translated them into Urdu to facilitate multilingual evaluation. To ensure both clinical validity and linguistic consistency, all generated and translated question-answer pairs were reviewed and validated by a clinical expert. The resulting **WBCMor_VQA** resource contains paired micro-

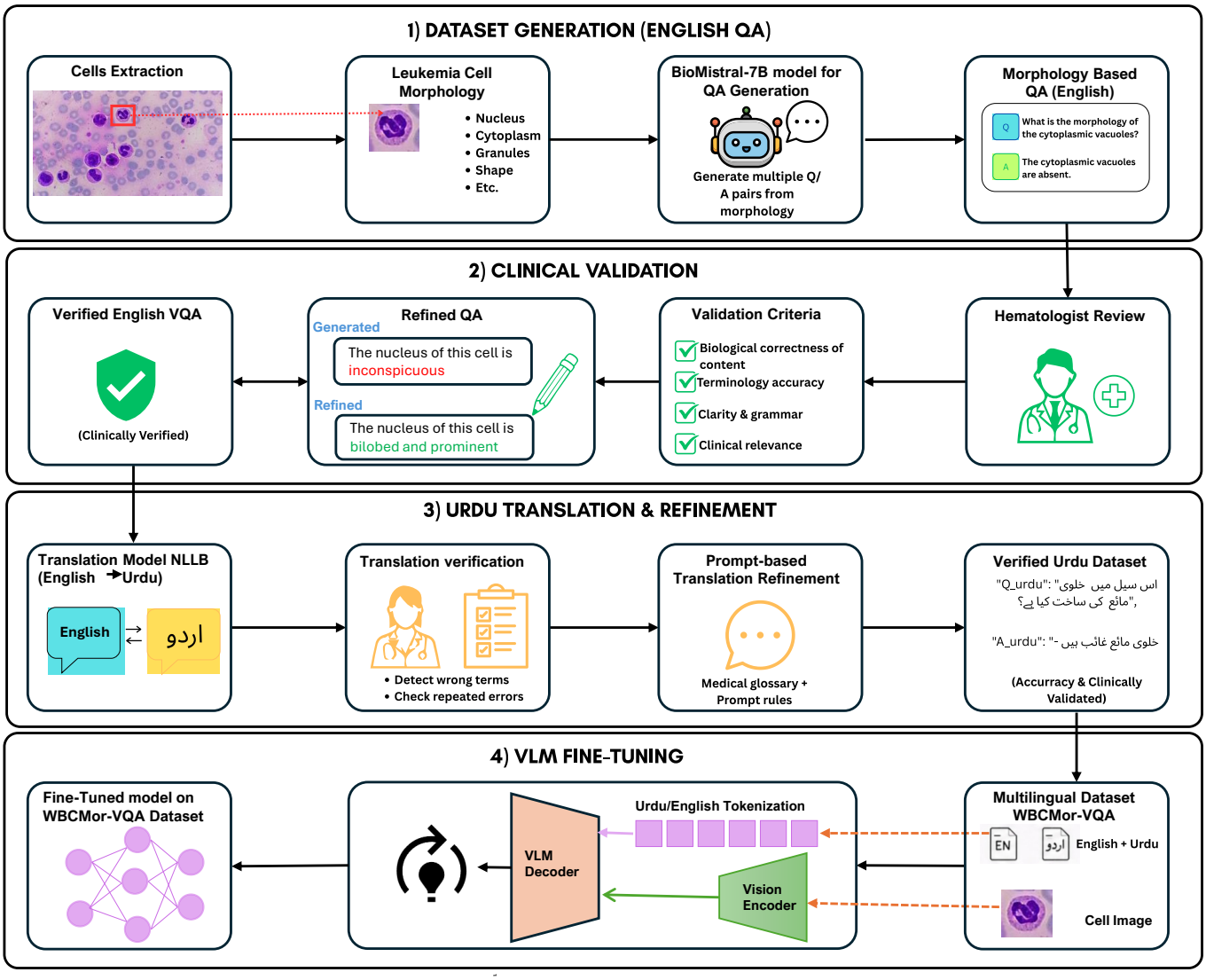


Figure 1: Pipeline of the proposed bilingual WBCMor-VQA dataset construction and model development pipeline. Cell-level morphology attributes are converted into English VQA pairs using BioMistral-7B (Labrak et al., 2024), followed by expert validation. The dataset is then translated into Urdu using NLLB (Costa-Jussà et al., 2022) and refined using a domain-specific medical dictionary and expert review. Finally, the resulting bilingual dataset is used to fine-tune vision-language models for hematology question answering.

scopic single-cell images with bilingual annotations in English and Urdu. By integrating morphology-aware visual information with multilingual QA representations, the dataset supports the development and evaluation of VLM's for understanding leukemic and normal white blood cell morphology, as well as medical AI applications.

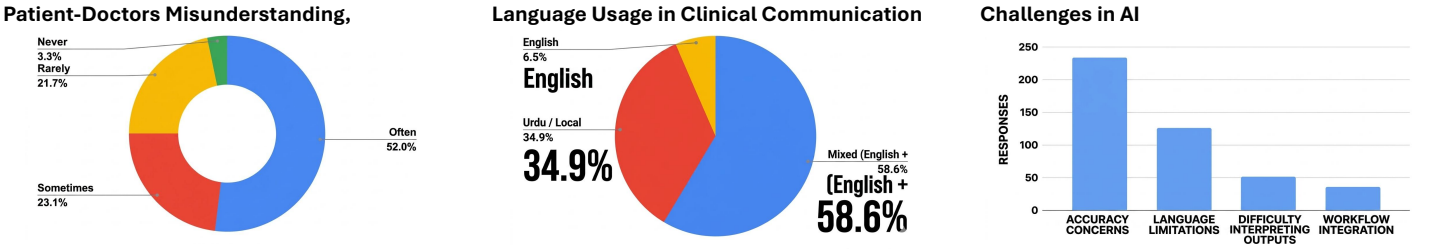
Our main contributions are as follows:

1: A cross-sectional survey involving physicians from public and private healthcare institutions was conducted to investigate language-related challenges in clinical communication and the adoption of AI-assisted medical tools. The survey explored language preferences, communication barriers, AI usage patterns, and clinicians' perspectives on improving patient understanding through AI-supported solutions.

2: We develop and clinically validate cell-level bilingual (English and Urdu) Visual Question Answering datasets for leukemia and normal white blood cells, comprising 110K question–answer pairs corresponding to single-cell images and enriched with morphological knowledge.

3: We construct a domain-specific Urdu hematology dictionary to standardize morphology-related terminology and ensure translation consistency across hematology QA pairs.

4: We evaluate multiple open-source Vision-Language Models on the proposed bilingual datasets and provide a comprehensive analysis of their performance on hematology visual question answering in both English and Urdu.



(a) (b) (c)

Figure 2: Survey findings related to physician communication practices and challenges in AI-assisted healthcare systems. (a) Frequency of patient misunderstanding of medical explanations, (b) language preferences used during clinical communication, and (c) challenges encountered while using AI-assisted healthcare systems.

3. Discussion

Effective medical communication remains a challenge in multilingual clinical environments, particularly in low- and middle-income countries such as Pakistan. To better understand these challenges and the growing role of artificial intelligence in healthcare, we conducted a survey involving practicing physicians across Pakistan. The findings provide empirical evidence of communication gaps in clinical practice and motivate the need for multilingual vision-language resources, supporting the development of our bilingual hematology VQA dataset.

3.1 Clinical Insights and Dataset Motivation

A cross-sectional survey was conducted to investigate communication practices, language barriers, and the adoption of AI-assisted tools in clinical settings. The survey involved 430 practicing physicians from public and private healthcare institutions across multiple regions of Pakistan. We have made the survey questionnaire available online [1].

The survey revealed substantial communication challenges in routine clinical practice, with 75.1% of respondents reporting that patients frequently or occasionally misunderstand medical explanations. Physicians commonly relied on mixed Urdu–English communication to improve patient understanding, reflecting the multilingual nature of clinical practice. Additionally, 73.2% of respondents reported using AI-assisted tools for diagnostic support, report summarization, clinical documentation, and medical education. Qualitative responses further highlighted the lack of standardized Urdu terminology in specialized medical domains, particularly pathology and hematology, with many physicians relying on English terms or informal translations to convey complex diagnostic concepts.

The distribution of physician responses to the question shown in Figure 2a, "Have you encountered situations where patients misunderstood medical explanations?", indicates that communication barriers are common in routine clinical practice. Overall, 52.0% of respondents said that misunderstandings occur often, 23.1% said they occur sometimes, 21.6% reported that they occur rarely, and only 3.3% said they never encounter such misunderstandings. Figure 4 presents the responses to the question, "What language do you primarily use to communicate with patients?" Most physicians (58.6%) reported using a mixed English–Urdu–local language communication style, followed by Urdu or local languages (34.9%), whereas only 6.5% relied exclusively on English. These findings emphasize the multilingual nature of clinical communication in Pakistan. The challenges encountered by clinicians when using AI-assisted healthcare systems are illustrated in Figure 2c. Accuracy concerns were reported most frequently, followed by workflow integration issues, difficulty interpreting AI-generated outputs, and language-related limitations. These observations suggest that concerns regarding reliability, interpretability, and integration continue to influence physician trust and adoption of AI-assisted systems. Collectively, these findings directly informed our dataset design. Multilingual communication motivated the inclusion of bilingual VQA pairs, while concerns about reliability and interpretability led to expert-validated annotations and medically grounded reasoning. Together, these considerations shaped the proposed English–Urdu hematology VQA dataset to support clinically relevant and linguistically accessible AI systems for medical education, decision support, and patient communication.

Limitations and Future Implications: While the survey provides valuable insights into clinical communication practices, it is limited to physicians practicing in Pakistan and therefore may not fully generalize to healthcare systems in other regions. In addition, the findings are based on self-reported responses and may be influenced by response and recall bias.

1. Link of the survey form: Clinical Language and AI Usage

Table 1: Comparison of hematology morphology and vision-language datasets

Dataset	Normal WBC	Leukemia WBC	Task Type	VQA	Bilingual	Expert Validated	Cell Images	QA Pairs
WBCAtt	✓	✗	Morphology Attributes	✗	✗	✓	10,298	–
LeukemiaAttri	✗	✓	Morphology Attributes	✗	✗	✓	∼10,000	–
WBCAtt-VQA (Uni-Hema)	✓	✗	VQA	✓	✗	✓	10,298	20,944
LeukemiaAttri-MLM (Uni-Hema)	✗	✓	Captioning/MLM	✗	✗	✓	∼10,000	-
HemBLIP	✓	✓	Image–Text	✗	✗	Automatic	14,659	-
Ours	✓	✓	**Bilingual VQA**	✓	✓	✓	**∼20,000**	**110,520**

4. Resource Availability

The dataset is publicly available on Figshare [2]. We also provide a demo on the Hugging Face platform [3]. The source code and pretrained models, along with documentation for fine-tuning, are available on GitHub [4].

Potential Use Cases: The published dataset can be used to train vision and large language models for research purposes only. Currently, we strongly discourage the use of the resulting models (or any model trained on the given dataset) in real clinical practice or medical decision-making.

Licensing: The dataset, source code, and pretrained models are released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Ethical Considerations: This work utilizes publicly available, clinically validated medical datasets and contains no patient-identifiable information. Following question-answer generation from the available annotations, the dataset was reviewed and validated by a medical professional. After translation, the dataset was further evaluated by two qualified native Urdu speakers, with the assistance of an English-Urdu medical dictionary, to ensure both clinical accuracy and linguistic quality.Despite these validation steps, a small number of translation inconsistencies may remain. Given the large scale of the dataset, we will continue to conduct systematic reviews, and any identified or reported inconsistencies will be carefully examined and corrected in subsequent dataset updates.

5. Methodology

This section presents the pipeline for constructing the proposed bilingual hematology VQA benchmark. We begin with a clinical survey conducted to identify communication challenges and language requirements in healthcare settings. Next, we describe the construction of morphology-aware VQA datasets using LeukemiaAttri (Rehman et al., 2024) and WBCAtt (Tsutsui et al., 2023), followed by expert validation, Urdu translation, translation quality assessment, terminology refinement, and prompt-guided translation. The overall workflow is illustrated in Figure 1. Finally, we outline the experimental setup and evaluation of multiple VQA models on the proposed benchmark.

5.1 Clinical Survey Design and Data Collection

To understand communication practices, language barriers, and the adoption of artificial intelligence in clinical workflows, we conducted a cross-sectional survey of practicing physicians across Pakistan. The survey was designed to capture clinical communication behavior in multilingual healthcare environments. The questionnaire was structured into four major sections: (1) demographic information, including professional role and institutional affiliation; (2) physician–patient communication practices, focusing on language usage and communication strategies; (3) use of AI tools in clinical workflows, including diagnostic support and documentation; and (4) perceptions of AI reliability, trust, and clinical responsibility. The survey instrument consisted of both multiple-choice and Likert-scale questions, along with optional open-ended responses to capture qualitative feedback. The questionnaire was developed in English and designed to be easily understood by clinical practitioners. Data collection was conducted using Google Forms, and participation was voluntary and anonymous. The survey was distributed through professional medical networks across multiple healthcare institutions in Pakistan, including both public and private-sector hospitals. A total of 430 physicians participated in the study. No personally identifiable information was collected, ensuring participant confidentiality and compliance with ethical research standards.

5.2 Leukemia Morphology VQA Construction

To develop a morphology-aware WBCMor-VQA benchmark for leukemic and normal cells, we utilized the publicly available Leukemia patient dataset (Rehman et al., 2025, 2024). The dataset provides cell-level morphology annotations describing the 13 WBC classes and six clinically relevant attributes of each WBC: nuclear chromatin, nuclear shape, nucleolus characteristics, cytoplasm appearance, cytoplasmic basophilia, cytoplasmic vacuoles, and cell size. Although these annotations provide detailed morphological information, the dataset does not contain the textual de-

2. Figshare: https://doi.org/10.6084/m9.figshare.32727159
3. Hugging Face: https://huggingface.co/spaces/ryhm/WBC-Mor-VQA
4. GitHub:https://github.com/intelligentMachines-ITU/WBC-Mor-VQA-dataset

Table 2: The table shows the total number of corrections made in each category together with representative examples. During validation, experts refined unclear questions, corrected incorrect answers, fixed grammar and formatting errors, and revised medically inaccurate descriptions of morphology.

Correction Category	Train Set	Test Set	Model Generated	Expert Validated
Question	09	84	What is the nuclear chromatin pattern of this leukemia cell?	What is the nuclear chromatin pattern of this cell?
Answer	1238	761	The nuclear shape of this cell is irregular.	The nuclear shape of this cell is round.
Grammar	38	285	The cytoplasm in thisneutrophilic cell is abundant.	The cytoplasm in this neutrophilic cell is abundant.
Medical	1209	560	The nucleus of this cell is inconspicuous.	The nucleus of this cell is bilobed and prominent.
Total Corrections	1247	845	–	–

tails. Therefore, we transformed the morphology annotations into structured VQA samples to enable morphology-oriented vision-language learning and reasoning.

Visual Question Answer Generation: We generated question answer pairs using BioMistral-7B (Labrak et al., 2024) following the VQA generation strategy adopted in UniHema (Rehman et al., 2026). For each cell image, the six morphology attributes were provided as structured input to the model. Based on these attributes, BioMistral-7B (Labrak et al., 2024) generated clinically relevant questions together with their corresponding answers. The generated questions focused on morphology-related characteristics commonly examined during hematological assessment, including chromatin patterns, nuclear morphology, cytoplasmic appearance, and cytoplasmic vacuoles. Depending on the available annotations, between two and six question-answer pairs were generated for each cell image.

Prompt Design: To generate morphology-oriented VQA pairs, morphology annotations associated with each leukemia cell were first converted into structured textual descriptions and provided as input to BioMistral-7B (Labrak et al., 2024). The model was prompted to act as a hematology expert and generate clinically relevant question–answer pairs grounded exclusively in the provided morphology attributes. The prompt encouraged diverse question formulations, morphological reasoning, and medically consistent responses while constraining the output to the given annotations. The prompt template used for VQA generation is shown in Figure 3.

Expert Validation of English VQA Samples: The source dataset was previously validated by domain experts. Additionally, the English dataset was reviewed by a medical expert [5], during which each question–answer pair was assessed alongside its corresponding cell image and morphology annotations for clinical relevance, clarity, and consistency. The expert verified that the questions were well-formed and clinically meaningful and that the answers accurately reflected the underlying morphological attributes. Any instances of ambiguity, grammatical issues, clinical irrelevance, or annotation mismatches were corrected. Table 2 summarizes the types of errors identified along with representative examples from the validation process.

You are creating a medical VQA dataset.
Generate multiple question–answer pairs from this morphology.

Rules:

- Use provided morphology information and your medical knowledge with respect to leukemia.
- Answers must be complete, detailed sentences.
- Do not repeat question templates.
- Questions may combine multiple morphology attributes.
- Return only valid JSON.

Morphology: `{morphology_text}` **Return format:**

```
{"q1":"...", "a1":"...", ...}
```

Figure 3: Prompt template used for morphology-based VQA generation with BioMistral-7B. Structured leukemia cell morphology attributes are provided as input, and the model generates multiple clinically grounded question-answer pairs in JSON format.

5.3 Urdu Dataset Construction

The validated English question-answer pairs were translated into Urdu using the NLLB-200 translation model (Costa-Jussà et al., 2022). Translation was performed using the `eng_Latn` to `urdu_Arab` language pair while preserving the original image-question-answer relationship.

Translation Error Analysis: After translation, several inconsistencies were identified in morphology-related terminology. The Urdu dataset was therefore manually reviewed using an English-to-Urdu medical dictionary with assistance

5. Dr. Hafiza Tooba Aftab (MBBS, M.Phil. Hematology, FCPS Resident, Hematology) independently reviewed all English VQA pairs for clinical accuracy and consistency.

Table 3: Examples of corrections in the Urdu dataset.

English Term	NLLB Translation	Corrected Urdu
Cell	سیل	خلیہ
Nucleus	نیوکلیئس	مرکزہ
Cytoplasm	سائٹوپلاسم,	خلوی مائع
Morphology	مورفولوجی	ساخت
Granules	گرانولز, گارنٹیڈ	دانے دار

from native Urdu speakers to detect translation errors in hematology-specific terms. It was observed that key domain concepts such as nucleus, chromatin, morphology, and other cellular attributes were sometimes translated into incorrect or clinically inappropriate Urdu terms. To resolve this issue, a domain-specific English-to-Urdu hematology dictionary was developed and applied as a post-processing step. This correction process resulted in more than five thousand terminology fixes across the dataset, and the revised translations were incorporated into the final Urdu benchmark. Table 3 summarizes these corrections.

Prompt-Guided Re-translation: After constructing the Urdu hematology dictionary, the translated dataset was reprocessed using a refined translation prompt. The translation model was explicitly instructed to preserve the standardized Urdu terminology defined in the dictionary whenever corresponding English medical terms appeared in the source text. This process reduced translation variability and improved consistency across morphology descriptions, questions, and answers, resulting in the final Urdu version of the leukemia morphology VQA dataset.

5.4 WBCAtt-VQA Bilingual Extension

In addition to the leukemia morphology VQA dataset, we extended the validated WBCAtt-VQA dataset introduced in Uni-Hema (Rehman et al., 2026). The dataset was processed using the same multilingual pipeline as the leukemia morphology dataset. Specifically, the English question–answer pairs were translated into Urdu using NLLB-200 (Costa-Jussà et al., 2022), followed by translation quality assessment, dictionary-guided terminology standardization, and prompt-guided re-translation. This process resulted in a bilingual version of WBCAtt-VQA containing aligned English and Urdu annotations.

5.5 Final Bilingual Benchmark

The final benchmark consists of two clinically validated, morphology-aware, bilingual white blood cell (WBC) VQA resources for leukemic and normal cells: the newly developed leukemia morphology (Rehman et al., 2026) VQA dataset and the bilingual extension of normal morphology WBCAtt-VQA (Rehman et al., 2026). Each sample contains a microscopic cell image together with semantically equivalent question-answer pairs in English and Urdu. For the English benchmark, the LeukemiaAttri dataset contains 31,029 training samples and 9,424 testing samples, while the WBCAtt dataset contains 12,339 training samples and 2,468 testing samples. For the Urdu benchmark, the translated and clinically validated LeukemiaAttri dataset contains 31,029 training samples and 9,424 testing samples, whereas the Urdu WBCAtt dataset contains 12,339 training samples and 2,468 testing samples.

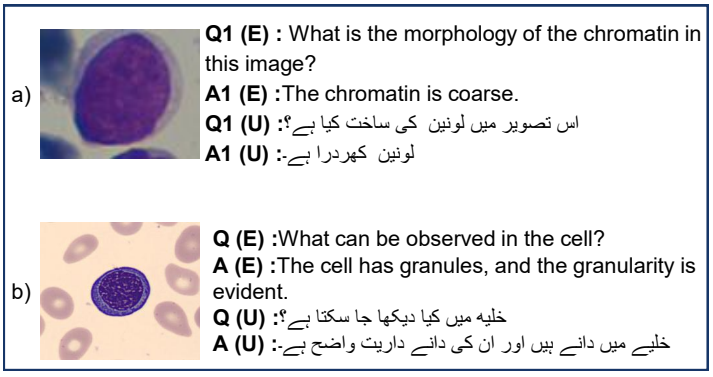


Figure 4: Examples of bilingual (English and Urdu) question–answer pairs from (a) leukemic cells and (b) normal white blood cells.

5.6 Model Fine-tuning and Evaluation Setup

To demonstrate the usability of the proposed hematology VQA resource, we evaluated multiple VLMs under zero-shot and fine-tuning settings. We evaluated two general-purpose VLMs, Qwen2-VL-2B-Instruct (Wang et al., 2024) and InternVL2.5-2B (Chen et al., 2024), along with a hematology specific unified model (Rehman et al., 2026), to assess the dataset's effectiveness for both general and biomedical VQA. Experiments were conducted on the WBCMor-VQA English, Urdu, and combined sets. For fine-tuning, each sample consisted of a microscopic cell image paired with a morphology-related question and its corresponding answer. Models were trained using Low-Rank Adaptation (LoRA) (Che et al., 2026), while full supervised fine-tuning was applied to Uni-Hema with small-T5 where applicable to improve efficiency and reduce memory overhead. The models were optimized to generate concise morphology-specific responses, including nuclear chromatin pattern, nuclear shape, cytoplasmic appearance, and basophilia level. The training and test splits were kept strictly separate to avoid data leakage. Evaluation was performed on both expert-validated and non-validated test sets to analyze the impact of annotation refinement on model performance.

Table 4: Performance comparison of VLM's on the English, Urdu, and bilingual (English+Urdu) benchmarks.

Model	Setting	LeukemiaAttri			WBCAtt			LeukemiaAttri + WBCAtt		
		BLEU-4	ROUGE-L	BERTScore	BLEU-4	ROUGE-L	BERTScore	BLEU-4	ROUGE-L	BERTScore
Zero-shot/Fine-tune and test on English										
Qwen2-VL-2B-Instruct	Zero-shot	0.232	0.409	0.909	0.090	0.286	0.878	0.203	0.384	0.903
Qwen2-VL-2B-Instruct	Fine-tuned	0.823	0.894	0.983	0.274	0.536	0.939	0.653	0.819	0.969
InternVL2.5-2B	Fine-tuned	0.149	0.261	0.887	0.052	0.260	0.886	0.100	0.261	0.887
Uni-Hema(T5-Small)	Fine-Tuned	0.659	0.836	–	**0.322**	**0.557**	–	0.589	0.778	–
Zero-shot/Fine-tune and test on Urdu										
Qwen2-VL-2B-Instruct	Zero-shot	0.049	–	0.911	0.043	–	0.910	0.047	–	0.910
Qwen2-VL-2B-Instruct	Fine-tuned	0.112	–	0.884	0.017	–	0.842	0.095	–	0.875
Fine-tune on English + Urdu and test on Urdu										
Qwen2-VL-2B-Instruct	Fine-tuned	0.291	–	0.953	0.121	–	0.932	0.257	–	0.947
Fine-tune on English + Urdu and test on English										
Qwen2-VL-2B-Instruct	Fine-tuned	**0.834**	**0.899**	**0.984**	0.274	0.521	0.932	**0.661**	**0.821**	**0.973**

Table 5: Qwen2-VL-2B-Instruct results on Non-Validated English, Urdu, and bilingual (English+Urdu) benchmarks.

Model	Setting	LeukemiaAttri			WBCAtt			LeukemiaAttri + WBCAtt		
		BLEU-4	ROUGE-L	BERTScore	BLEU-4	ROUGE-L	BERTScore	BLEU-4	ROUGE-L	BERTScore
Fine-tune on English + Urdu and test on English (Non-validate)										
Qwen2-VL-2B-Instruct	Fine-tuned	0.839	0.898	0.982	0.285	0.516	0.921	0.667	0.822	0.970
Fine-tune on English + Urdu and test on Urdu (Non-Validate)										
Qwen2-VL-2B-Instruct	Fine-tuned	0.017	-	0.846	0.002	-	0.8328	0.009	-	0.841

6. Validation

To validate the proposed English–Urdu hematology VQA dataset, we conduct a comprehensive evaluation using vision–language models on English and Urdu benchmarks. Strict non-overlapping train/test splits are maintained to ensure unbiased assessment and robust generalization analysis. The evaluation is performed in three settings: English-only, Urdu-only, and a combined bilingual benchmark, in which the same model is evaluated directly across both languages to assess its cross-lingual generalization capability. Model outputs are assessed using BLEU-4, ROUGE-L, and BERTScore, providing complementary lexical and semantic measures of response quality.

6.1 Implementation Details

For Qwen2-VL-2B-Instruct, the maximum sequence length was set to 512 tokens, and the model was trained for three epochs with a learning rate of 5×10^{-5}, using gradient accumulation to handle limited GPU memory. For Uni-Hema, we used a pretrained T5-small model, while the rest of the network was trained from scratch using standard settings. For InternVL2.5-2B, images were resized to 448×448 pixels, and the model was fine-tuned using LoRA-based supervised learning. The sequence length was set to 768 tokens, with a learning rate of 1×10^{-5} over three epochs, along with gradient accumulation for stable optimization. During inference, greedy decoding was used with a maximum generation length of 32 tokens. Evaluation was performed using BLEU-4 (Papineni et al., 2002), ROUGE-L, and BERTScore F1 (Liu et al., 2019) to assess lexical accuracy, performance, and semantic similarity.

6.2 Experiments and Results

The performance comparison across the English, Urdu, and bilingual (English+Urdu) benchmarks is presented in Table 4. The results demonstrate that domain-specific fine-tuning consistently outperforms zero-shot evaluation, highlighting the effectiveness of the proposed bilingual hematology VQA dataset. Furthermore, the observed trends indicate that leveraging additional bilingual supervision enables models to generalize across both languages, suggesting that similar performance gains may be achieved when such cross-lingual resources are used to train other vision-language models. Although Table 5 shows that the English non-validated benchmark achieves slightly higher BLEU-4 score, expert-validated annotations provide a more reliable basis for model evaluation by reducing annotation noise and improving benchmark quality (Wei et al., 2022; Bernhardt et al., 2022). In contrast, the non-validated Urdu version achieves substantially lower performance, whereas the Urdu dataset underwent extensive refinement, including large-scale terminology correction, prompt-guided re-translation, and expert validation.

7. Conflicts of Interest

The authors have no conflicts of interest to declare.

8. Acknowledgments

We would like to express our sincere gratitude to Saadia Malik, Zahra Malik, and Fatima Athar Khan for their continuous support, encouragement, and valuable contributions throughout this research. We also acknowledge Motives for providing the computational resources that made this research possible.

References

Farhana Badar and Shahid Mahmood. Cancer in lahore, pakistan, 2010–2019: an incidence study. *BMJ open*, 11 (8):e047049, 2023.

Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. Vision–language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380, 2023.

Mélanie Bernhardt, Daniel C. Castro, Ryutaro Tanno, Anton Schwaighofer, Kerem C. Tezcan, Miguel Monteiro, Shruthi Bannur, Matthew P. Lungren, Aditya Nori, Ben Glocker, Javier Alvarez-Valle, and Ozan Oktay. Active label cleaning for improved dataset quality under resource constraints. *Nature Communications*, 13(1), March 2022. ISSN 2041-1723. . URL `http://dx.doi.org/10.1038/s41467-022-28818-3`.

Kamlesh M Bhojwani, Ahmed Raheem, Urooba Tariq Khan, Fahad Javid, Daniyal Tanweer, Nawal Rehmani, Nadeem Ullah Khan, Saqib Raza Khan, et al. Trends in cancer burden in pakistan: a 30-year analysis from the global burden of disease study (1990–2019). *ecancer2026*, 2026.

Blood Cancer United. Blood cancer facts and statistics, 2025. URL `https://bloodcancerunited.org/blood-cancer/blood-cancer-facts-and-statistics`.

Lyle Campbell and Grondona. Ethnologue: Languages of the world. *Language*, 84(3):636–641, 2008.

Chang Che, Ziqi Wang, Pengwan Yang, Cheems Wang, Hui Ma, and Zenglin Shi. Lora in lora: Towards parameter-efficient architecture expansion for continual visual instructiontuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 19978–19986, 2026.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Adeline De Angelis. The ef english proficiency index as an international large-scale assessment: a critical cultural political economy perspective. *Globalisation, Societies and Education*, 23(2):478–491, 2025.

Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7: 1430984, 2024.

Béria Chingnabé Kalpélbé, Angel Gabriel Adaambiik, and Wei Peng. Vision language models in medicine. *arXiv preprint arXiv:2503.01863*, 2025.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pages 5848–5864, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL `https://arxiv.org/abs/1907.11692`.

Maimuna Biswas Noshin, Monoronjon Dutta, Md Nadim Kaysar, Rakib Hossain Sajib, Md Jakir Hossen, Dip Nandi, Abdullah Al Jubair, and Mashiour Rahman. Medical visual question answering with multimodal: a systematic mini review (2023–2026). *Frontiers in Digital Health*, 8:1848710, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318. Association for Computational Linguistics, 2002.

S. Poria et al. Linguistic diversity and language access in south asian healthcare. *South Asian Social Review*, 2025.

Abdul Rehman, Talha Meraj, Aiman Mahmood Minhas, Ayisha Imran, Mohsen Ali, and Waqas Sultani. A large-scale multi domain leukemia dataset for the white blood

cells detection with morphological attributes for explainability. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 553–563. Springer, 2024.

Abdul Rehman, Talha Meraj, Aiman Mahmood Minhas, Ayisha Imran, Mohsen Ali, Waqas Sultani, and Mubarak Shah. Leveraging sparse annotations for leukemia diagnosis on the large leukemia dataset. *Medical Image Analysis*, page 103760, 2025.

Abdul Rehman, Iqra Rasool, Ayisha Imran, Mohsen Ali, and Waqas Sultani. Uni-hema: Unified model for digital hematopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 37578–37589, 2026.

Satoshi Tsutsui, Winnie Pang, and Bihan Wen. Wbcatt: A white blood cell dataset annotated with detailed morphological attributes. *Advances in Neural Information Processing Systems*, 36:50796–50824, 2023.

Muhammad Tufail and Changxin Wu. Cancer statistics in pakistan from 1994 to 2021: data from cancer registry. *JCO Clinical Cancer Informatics*, 7:e2200142, 2023.

Julie van Logtestijn and Petru Manescu. Hemblip: A vision-language model for interpretable leukemia cell morphology analysis. *arXiv preprint arXiv:2601.03915*, 2026.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations, 2022. URL `https://arxiv.org/abs/2110.12088`.