NAKBA NLP 2026: Shared Task on Arabic Handwritten Manuscript Understanding (Palestine Memory–Omar Al-Saleh Memoir)

Hadi Hamoud[1], Ahmad Ali Chamseddine[2], Bilal Shalash[1], Firas Ben Abid[3], Mustafa Jarrar[4], Chadi Abou Chakra[1], Bernard Ghanem[5], Fadi A. Zaraket[1]
[1]Arab Center for Research and Policy Studies, Doha, Qatar
[2]Doha Institute, Doha, Qatar
[3]Zinki AI
[4]Hamad Bin Khalifa University, Doha, Qatar
[5]King Abdullah University of Science and Technology (KAUST), Saudi Arabia
{hhamoud, bshalash, cabouchakra, fzaraket}@dohainstitute.edu.qa
achamsed@dohainstitute.edu.qa, firas.ben.abid@zinki.ai
mjarrar@hbku.edu.qa, bernard.ghanem@kaust.edu.sa

Abstract

Transcribing historical Arabic manuscripts into machine-readable text is essential for preserving cultural heritage and enabling computational research in the humanities, yet it remains a challenging task due to handwriting variability, page degradation, and the complexity of Arabic script. To advance research in this area, we introduce the NAKBA NLP 2026 shared task on Arabic manuscript understanding, comprising two complementary tracks: a manual transcription track, in which participating teams annotate unlabelled handwritten line images, and an automatic system track for handwritten text recognition (HTR). Both tracks use the Omar Al-Saleh Memoir Collection, a corpus of 6,395 scanned pages and approximately 1.6 million words, written between 1951 and 1965 and provided by the Palestine Memory Project. The dataset, evaluation scripts, and system outputs are publicly available. In Subtask 1 (Transcription Track), three teams contributed manual line-level transcriptions; evaluation on hidden ground-truth samples yielded Character Error Rates (CER) between 0.06 and 0.11. In Subtask 2 (Systems Track), seven teams submitted HTR systems. The top-performing system, by Misraj AI, achieved a corpus-level CER of 0.079 and Word Error Rate (WER) of 0.244, outperforming the organiser baseline (CER 0.368, WER 0.691). Rankings shift between corpus-level and per-line evaluation: the Ketaba OCR team achieved the lowest per-line CER (0.082). All contributed transcriptions and system outputs are released under CC-BY-4.0 to support continued research in Arabic manuscript recognition and digital humanities.

Keywords: Arabic manuscripts, handwritten text recognition, shared task, dataset, digital humanities

1. Introduction

Arabic historical manuscripts constitute a central source for research in history, politics, and the digital humanities, providing first-hand accounts of social, intellectual, and political life across the modern Arab world (Alrobah and Alzahrani, 2022; Saeed et al., 2024). However, much of this material remains inaccessible to computational analysis, as it is preserved primarily in handwritten form and lacks structured digital representations. Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) offer pathways to large-scale digitisation and search over such collections. However, Arabic Manuscript Handwritten Text Recognition (AMHTR) remains challenging due to handwriting variability, page degradation, complex layouts, and Arabic-specific natural language processing requirements such as short vowel (diacritic) omission and rich morphology (Graves et al., 2006; Mahmoud et al., 2014).

Progress in AMHTR is currently limited by two key gaps. First, publicly available, well-annotated datasets that reflect the characteristics of historical Arabic handwriting—including diversity of writing styles, ink quality, and page structure—remain scarce (Saeed et al., 2024). Second, the NLP community lacks established benchmarks and competitive evaluations targeting AMHTR specifically. This stands in contrast to the well-developed ecosystem of shared tasks for Latin-script historical document recognition, such as the ICDAR HTR competitions (Romero et al., 2012) and the RASM series (Clausner et al., 2018).

Recent advances in vision–language models and document analysis pipelines suggest that modern approaches, when combined with layout segmentation and domain-specific annotation, may yield viable solutions for manuscript recognition even in low-resource settings (Li et al., 2023; Bai et al., 2023). However, realising this potential requires open datasets, standardised evaluation protocols, and community engagement through competitive shared tasks.

To address these gaps, we introduce NAKBA NLP 2026, a shared task for Arabic manuscript understanding organised as part of the second International Workshop on Nakba Narratives as Lan-

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 70–79

guage Resources (Jarrar et al., 2026). The shared task is built on the Omar Al-Saleh Memoir Collection (Institute for Palestine Studies, a; Arab Center for Research and Policy Studies, 2024), which comprises 6,395 scanned manuscript pages written between 1951 and 1965, with line-level annotations for a curated subset. The shared task is organised into two complementary tracks. **Subtask 1 (Transcription Track)** invites teams to produce expert-quality manual transcriptions of unlabelled handwritten line images, enriching the available ground truth for future research. **Subtask 2 (Systems Track)** challenges teams to develop automatic AMHTR systems, evaluated using Character Error Rate (CER) and Word Error Rate (WER).

The main contributions of this shared task are: (1) the public release of a large-scale, line-level annotated Arabic manuscript dataset with standardised train, development, and test splits; (2) a dual-track evaluation framework that combines manual transcription quality assessment with automatic system benchmarking; and (3) an empirical analysis of system performance, error patterns, and the relationship between corpus-level and per-line evaluation metrics.

In Subtask 1, three teams contributed manual transcriptions, achieving CER between 0.06 and 0.11 on hidden ground-truth samples, demonstrating both the feasibility and the difficulty of expert Arabic manuscript transcription. In Subtask 2, all seven participating systems outperformed the baseline. The top-performing system, by Misraj AI, reached a corpus-level CER of 0.079 and WER of 0.244, compared with 0.368 and 0.691 for the baseline. Rankings differ between corpus-level and per-line evaluation: Ketaba OCR obtained the lowest per-line CER (0.082), while Misraj AI led on corpus-level CER.

The remainder of this paper is organised as follows. Section 2 reviews related work. Section 3 describes the dataset. Section 4 presents the shared task design. Section 5 details the baseline system. Section 6 summarises participating systems. Section 7 reports results and analysis. Section 8 concludes with future directions.

2. Related Work

This section situates the present shared task within three areas of related work: existing AMHTR datasets, OCR and HTR model architectures, and prior shared tasks in document recognition. We highlight the gaps that motivated the design of NAKBA NLP 2026.

AMHTR Datasets. Several datasets have been developed for Arabic handwriting recognition, though they vary considerably in scale, annotation granularity, and public availability. The KHATT dataset (Mahmoud et al., 2014) provides 1,000 handwritten Arabic paragraphs from 1,000 writers, offering writer diversity but limited historical manuscript coverage. IFN/ENIT (Pechwitz et al., 2002) targets handwritten Tunisian city names, restricting its scope to isolated word recognition. The MADCAT corpus (Cieri et al., 2016), developed under the DARPA programme, offers annotated Arabic handwriting but imposes restrictive licensing that limits its use in open benchmarks. More recently, the Muharaf dataset (Saeed et al., 2024) has become the largest publicly available AMHTR collection, with over 36,000 line images from 1,600 historical manuscript pages spanning diverse document types. The VML-HD dataset (Kassis et al., 2017) provides historical document images, and recent surveys (Alrobah and Alzahrani, 2022) catalogue additional resources.

The Omar Al-Saleh Memoir dataset presented in this task differs from these prior resources in two respects. First, it provides large-scale, multi-year manuscript data from a single author with line-level annotations aligned to high-resolution page images, enabling the study of temporal variation in handwriting within a longitudinal collection. Second, with approximately 6,395 pages and 1.6 million words, it is substantially larger than most existing AMHTR corpora.

OCR and HTR Models. The trajectory of Arabic HTR has moved from traditional Hidden Markov Models and recurrent neural networks with Connectionist Temporal Classification (CTC) decoding (Graves et al., 2006) towards transformer-based architectures. TrOCR (Li et al., 2023) combines a vision transformer encoder with a language model decoder and has been adapted for Arabic script. Large vision–language models such as Qwen-VL (Bai et al., 2023), InternVL (Chen et al., 2024), and Florence-2 (Xu et al., 2024) have demonstrated multilingual OCR capabilities, though their application to historical Arabic manuscripts remains underexplored. These models provide the architectural foundation for several systems submitted to the present shared task, as described in Section 6.

Shared Tasks in Document Recognition. The ICDAR competitions on handwritten text recognition have established benchmarks for Latin, German, and Chinese scripts (Romero et al., 2012) but have not included dedicated Arabic manuscript tracks. The ICDAR Arabic Handwriting Recognition Competition series (2005–2011) focused on word-level recognition of modern handwritten town names using the IFN/ENIT database, rather than full-line transcription from historical documents. The RASM competitions (RASM2018 at ICFHR and

RASM2019 at ICDAR) targeted page-level recognition of historical Arabic scientific manuscripts from the British Library/Qatar Digital Library, evaluating both layout analysis and text recognition in PAGE format (Clausner et al., 2018). The OpenITI project (Romanov et al., 2023) has contributed OCR tools for classical Arabic texts, primarily targeting printed rather than handwritten material.

NAKBA NLP 2026 complements these efforts by providing a line-level AMHTR benchmark focused on single-author memoir transcription with pre-segmented line images and standardised CER/WER evaluation.

Crowdsourced and Manual Transcription. Manual transcription campaigns, such as those organised by the Transkribus platform (Muehlberger et al., 2019) and the READ project, have demonstrated the value of combining human expertise with computational tools for historical document analysis. Subtask 1 builds on this tradition by organising competitive manual transcription with standardised evaluation, providing a model for future Arabic manuscript annotation efforts.

3. Dataset

The dataset used in the shared task is derived from the Omar Al-Saleh Memoir Collection, a corpus of handwritten Arabic manuscripts written between 1951 and 1965. Omar Al-Saleh Al-Barghouti (1894–1965) was a Palestinian-Jordanian politician, lawyer, and author. Born in Deir Ghassaneh near Ramallah, he was active in the Palestinian national movement from 1919, served as a member of the Jordanian Parliament after the union of the two banks, held the position of Minister of Education in 1955 and again in 1959, and co-authored *Tarikh Filastin* (History of Palestine) with Khalil Totah (Al-Barghouti and Totah, 1923; Institute for Palestine Studies, a). His memoirs offer a first-person account of political, social, and personal life during a formative period of modern Palestinian and Arab history.

The dataset is provided by the Arab Center for Research and Policy Studies[1] as part of their work on the Palestine Memory Project[2] (Arab Center for Research and Policy Studies, 2024). The memoir collection is also archived by the Institute for Palestine Studies as part of the Palestine Social History Archives (Institute for Palestine Studies, b). It serves as an initial component in a broader sequence of curated Arabic manuscript corpora belonging to historical authors and figures.

[1] https://acr.ps

[2] https://palestine-memory.org/

3.1. Corpus Statistics

Table 1 provides per-document statistics. The full dataset comprises 6,395 scanned pages. After transcription and aggregation, it contains 1,597,025 words distributed across 50,685 sentences, with 161,394 unique word types and an overall type–token ratio of 0.1011. The average sentence length is 31.5 words. Substantial variation in document length is observed: the largest document (1959) contains 132,858 words across 532 pages, while the smallest (1957b) contains 39,985 words across 160 pages, reflecting differences in the author's narrative style and memoir writing activity across years.

Word and sentence counts were computed automatically using a Python-based text analysis pipeline. Words were extracted using regular expressions matching Arabic and Latin character sequences (single characters excluded), and sentences were identified by splitting on sentence-ending punctuation marks, including the Arabic question mark (؟) and the standard period, exclamation mark, and question mark.

Documents labelled with a "b" suffix (1957b, 1961b) represent separate notebooks written by the same author in the same year. These are treated as distinct documents to preserve the physical organisation of the collection.

Doc.	Pages	Words	Sent.
1951	475	118,742	2,470
1953	451	112,745	2,968
1954	454	113,482	3,105
1955	421	105,192	2,936
1956	328	81,780	2,197
1957	470	117,336	3,824
1957b	160	39,985	1,231
1958	342	85,341	2,597
1959	532	132,858	4,010
1960	526	131,363	4,594
1961	463	115,569	4,156
1961b	242	60,401	3,052
1962	466	116,257	3,846
1963	353	88,181	3,014
1964	504	125,837	4,689
1965	208	51,956	1,996
Total	**6,395**	**1,597,025**	**50,685**

Table 1: Omar Al-Saleh Memoir statistics. Documents with a "b" suffix denote separate notebooks from the same year.

3.2. Physical Characteristics

The manuscripts are digital photocopies of handwritten notebooks, with each scanned image typically containing two facing pages. Images are stored in JPEG format at a resolution of 1982×1400

pixels with a nominal resolution of 96 DPI. The manuscripts exhibit several features that make automated processing challenging: mixed layouts with headers, body text, and marginal notes; non-uniform line spacing that varies both within and across pages; handwriting variability in letter forms, ligatures, and word spacing; occasional overwritten or crossed-out words; and rare non-Arabic elements such as Western numerals or Latin-script annotations.

3.3. Annotation and Processing Pipeline

Layout Segmentation. Because each scanned image typically contains two facing manuscript pages, layout segmentation is applied first to separate the text into distinct regions. A YOLO11-based object detection model (Jocher et al., 2024; Khanam and Hussain, 2024), fine-tuned on manuscript page images, is used to detect the primary text regions, which are organised into four areas: left header, left body, right header, and right body. Detection accuracy was validated against manually annotated bounding boxes on a sample of 100 pages, achieving an Intersection over Union (IoU) above 0.92 for body regions. IoU measures the overlap between the predicted and ground-truth bounding boxes; a value of 0.92 indicates close alignment between detected and actual text areas. This metric applies to layout detection only and does not reflect downstream transcription accuracy.

Line Segmentation. To isolate individual text lines within each detected region, manuscript-specific line segmentation models are applied independently to each area. These models produce pixel-level masks for each text line. The masks are then converted into polygonal annotations that define the spatial extent of each line, and these polygons are used to extract cropped line images. The resulting line images serve as input for both manual transcription and automatic HTR. No automatic transcription system was used during the annotation process; all transcriptions were produced manually.

Manual Annotation. For each of the 16 manuscript documents, the first 15 pages are annotated at the line level, yielding 240 annotated pages and approximately 41,000 line-level annotations. Three trained annotators each transcribed an assigned portion of the data following standardised orthographic conventions covering normalisation rules, punctuation handling, and diacritics. Each annotator's transcriptions were then reviewed by the other two annotators. During this cross-review phase, normalised edit distance was computed between each original transcription and the reviewer's corrections; lines where the edit distance exceeded a predefined threshold were flagged for discussion and adjudication among the annotators. This procedure ensured consistency without requiring redundant independent annotation of the same lines. An additional 9,009 lines were subsequently manually annotated to enlarge the test set.

3.4. Data Splits

The annotated data is divided into training, development, and test sets as follows:

- **Training set**: 15,962 line images with gold transcriptions.
- **Development set**: 1,774 line images with gold transcriptions.
- **Test set**: 2,671 line images (transcriptions held out for evaluation).

3.5. Dataset Structure and Format

The dataset is organised hierarchically by document, page, region, and line. Each page is associated with metadata describing document origin, year, image dimensions, and filename. All annotations are stored in a JSON-based format preserving geometric information and reading order. For example, a single line annotation includes the document year, page number, region identifier (e.g., `left_body`), line number, polygonal coordinates defining the line boundary, and the corresponding transcription text. For Subtask 2, the data is also provided in CSV format, with each row containing the cropped line image filename and the gold transcription (empty for the test set).

4. Shared Task Design

4.1. Subtask 1: Transcription Track

Objective. This track focuses on producing expert-quality, line-by-line manual transcriptions of unlabelled manuscript line images, enriching the benchmark with reliable ground truth for training and evaluating HTR systems.

Data. Each team receives one mandatory batch of approximately 500 cropped line images from the Omar Al-Saleh Memoir Collection, together with an `annotations.csv` template containing columns for filename, text (to be filled), source page image reference, year, page number, and line number. In addition to the unlabelled images, each batch includes 50 pre-labelled samples drawn from the existing training dataset. These samples are embedded among the unlabelled images and serve

as hidden quality checkpoints: participants are not informed which samples carry ground-truth annotations, and evaluation is performed on these hidden samples. The full unannotated page images are also provided so that annotators can view the surrounding context when transcribing individual cropped lines. Teams that wish to contribute additional transcriptions may request access to further batches.

Requirements. Transcriptions must be manual and line-aligned (one transcription per image). Participants must follow the provided orthographic conventions covering normalisation rules, punctuation handling, and diacritics. No generative AI tools may be used to generate or correct transcriptions, whether fully or partially. Each team must submit an annotation guidelines document describing their procedures, handling of ambiguous or damaged text, consistency measures, and any ethical considerations.

Evaluation. Transcription quality is assessed along two dimensions: coverage and accuracy. *Coverage* is measured as the number of lines with valid Arabic transcriptions out of the 500 assigned. A transcription is considered valid if it contains at least one Arabic character and is not composed solely of non-Arabic content such as dashes, Western numerals, or Latin text. *Accuracy* is computed exclusively on the hidden pre-labelled samples for which verified ground truth is available (up to 50 per batch), using CER and WER as defined in Section 4.2. Only hidden samples that received a valid Arabic transcription contribute to the accuracy computation.

4.2. Subtask 2: Systems Track

Objective. This track requires participants to develop automatic systems that transcribe Arabic manuscript line images into machine-readable text. Participants may employ vision-based models, sequence-to-sequence architectures, or multimodal frameworks, and may fine-tune on the provided data or explore zero- and few-shot strategies.

Data. Participants are provided with the Omar Al-Saleh Memoir data using the splits described in Section 3. Each split contains cropped line images in JPEG format and an `annotations.csv` with image filenames and transcriptions (empty for the test set). The full unannotated page images are also available to support approaches such as page-level context modelling or self-supervised learning.

Evaluation Metrics. System performance is evaluated using Character Error Rate (CER) and Word Error Rate (WER), computed via Levenshtein edit distance, where edits comprise substitutions, insertions, and deletions. These metrics are reported in two complementary ways.

First, a *corpus-level* score sums all edit operations across the test set and divides by the total number of reference units (characters for CER, words for WER), as defined in Equations 3 and 4. Second, a *per-line* score computes the metric independently for each test line (Equations 1 and 2) and averages across all lines. Both variants are clipped to $[0, 1]$ and computed on the held-out test set of 2,671 lines. All results throughout this paper are reported on the $[0, 1]$ scale.

Formally, for a reference string r and hypothesis h, where d_{char} and d_{word} denote character-level and word-level edit distances respectively:

$$\text{CER}(r, h) = \frac{d_{\text{char}}(r, h)}{|r|} \tag{1}$$

$$\text{WER}(r, h) = \frac{d_{\text{word}}(r, h)}{|\text{split}(r)|} \tag{2}$$

At corpus level, if R_i and H_i denote the reference and hypothesis for the i-th test sample:

$$\text{CER}_{\text{c}} = \frac{\sum_i d_{\text{char}}(R_i, H_i)}{\sum_i |R_i|} \tag{3}$$

$$\text{WER}_{\text{c}} = \frac{\sum_i d_{\text{word}}(R_i, H_i)}{\sum_i |\text{split}(R_i)|} \tag{4}$$

Per-line variants are the arithmetic means of sample-level CER and WER over all evaluated lines. CER serves as the primary ranking metric, with corpus-level scores used for the official leaderboard.

5. Baseline System

The baseline system uses Qwen3-VL-8B-Instruct (Bai et al., 2023), a large open-weight vision–language model with demonstrated multilingual and OCR capabilities. We selected this model as a representative baseline because it allows us to measure the benefit of task-specific adaptation: the pretrained model has general document understanding ability but has not been trained on Arabic historical manuscripts.

The baseline approach consists of two steps. First, the pretrained model is adapted to the manuscript domain via parameter-efficient fine-tuning with LoRA, targeting the attention and feed-forward modules. This keeps the approach compute-efficient while preserving the pretrained representations. Second, the adapted model is used to generate transcriptions: given a cropped manuscript line image and a transcription instruction, the model produces the text directly. We de-

Component	Setting
Backbone	Qwen3-VL-8B-Instruct
Fine-tuning method	LoRA
LoRA rank / α / dropout	32 / 64 / 0.05
LoRA target modules	q,k,v,o,gate,up,down
Precision	bfloat16
Attention mechanism	FlashAttention-2
Epochs	2
Learning rate	2×10^{-5} (cosine)
Batch size / grad. accum.	24 / 1
Weight decay	0.01
Warmup ratio	0.03
Max seq. length	256 tokens
Max image resolution	1536 px (longest side)
Decoding	greedy, max 512 tokens

Table 2: Baseline configuration. The model is fine-tuned with LoRA on the shared-task training data.

code greedily to ensure deterministic, reproducible outputs.

During training, each example is formatted as a dialogue: a user turn containing the line image and instruction, followed by an assistant turn containing the gold transcription. The training objective is next-token prediction on the assistant response, with prompt tokens excluded from the loss. Images are resized with aspect ratio preservation. The same instruction prompt is used at both training and inference time.

Table 2 lists the full configuration. The baseline achieves a corpus-level CER of 0.368 and WER of 0.691 on the test set (2,671 lines), establishing a reference point for evaluating the benefit of more specialised system development.

6. Participating Systems

Table 3 summarises all participating teams, their affiliations, and the subtask(s) they entered.

Team	Affiliation	Subtask
PalNLP	Cardiff University	1
Sard	Arab Open University	1
Independent	Istanbul University	1
Misraj AI	Misraj AI	2
Oblevit	Arab Open University	2
Ketaba OCR	[to be confirmed]	2
Latent Narratives	Arab Open University	2
Al-Warraq	Arab Open University	2
Not Gemma	Helwan University	2
Fahras	Thakaa	2

Table 3: Participating teams, affiliations, and subtask(s).

6.1. Subtask 1 Participants

Three teams submitted transcriptions for Subtask 1. Teams varied in their handling of ambiguous or damaged text, diacritics, and lines containing non-Arabic content such as dates in Western numerals or standalone digits.

6.2. Subtask 2 Participants

Seven teams submitted systems for Subtask 2. The systems can be grouped by their underlying model families. Four teams built their systems around large vision–language models, primarily Qwen2-VL and InternVL2, adapted to the manuscript domain via LoRA fine-tuning. Three teams used encoder–decoder architectures based on TrOCR, with Arabic-specific tokenisation and data augmentation. Two teams additionally employed ensemble strategies that combined outputs from multiple models using character- or word-level voting. Data augmentation was common among competitive submissions, with elastic distortion, random erosion, contrast jittering, and additive noise being the most frequently used techniques. Three teams incorporated external Arabic text corpora for language model rescoring or decoder pre-training.

7. Results and Analysis

7.1. Subtask 1: Transcription Track Results

Table 4 summarises the coverage and accuracy results for all three participating teams. Coverage is measured as the number of lines receiving valid Arabic transcriptions out of 500, while CER and WER are computed only on the hidden pre-labelled samples that received valid annotations.

Team	Transcr.	Compl.	Test N	CER	WER
PalNLP	499	99.8%	49/50	**0.061**	**0.285**
Independent	497	99.4%	47/50	0.086	0.366
Sard	401	80.0%	39/50	0.114	0.433

Table 4: Subtask 1 results. Transcr. = lines with valid Arabic text. Compl. = completion rate. Test N = hidden pre-labelled samples with valid annotations out of 50.

PalNLP achieved the highest transcription quality with a CER of 0.061 and near-complete coverage (499 out of 500 lines). Independent Arabic Manuscript Transcription demonstrated comparable coverage (497 out of 500) and a CER of 0.086. Sard completed 401 out of 500 lines (80.0%), with 97 entries consisting solely of a dash character and 2 left blank; its CER of 0.114 on the annotated lines

indicates reasonable quality where transcriptions were provided.

Analysis. Across all teams, CER ranged from 0.061 to 0.114, confirming that manual transcription of Arabic manuscripts remains difficult even for trained human annotators, particularly for lines with degraded ink, ambiguous characters, or non-standard orthography. The variation in coverage highlights the need for clearer guidelines on handling non-Arabic content within manuscript pages.

Because each team transcribed a different batch, formal inter-annotator agreement scores (e.g., Cohen's kappa) cannot be computed across teams. Within the annotation pipeline (Section 3), cross-review among the three annotators and edit-distance-based flagging provided a quality control mechanism. Future iterations could assign overlapping subsets to multiple teams to enable direct comparison of annotator consistency.

The most common error patterns in manual transcriptions include confusion between visually similar characters such as *dal*/*dhal* (د/ذ), *sin*/*shin* (س/ش), and *ha*/*ta marbuta* (ه/ة); omission or misplacement of diacritical dots; and inconsistent handling of word boundaries in passages with minimal inter-word spacing.

7.2. Subtask 2: Corpus-Level Results

Table 5 reports the corpus-level CER and WER for all seven teams and the baseline on the held-out test set of 2,671 lines. CER is the primary ranking metric.

#	Team	CER	WER
1	Misraj AI	**0.079**	**0.244**
2	Oblevit	0.093	0.327
3	Ketaba OCR	0.094	0.300
4	Latent Narratives	0.105	0.311
5	Al-Warraq	0.114	0.378
6	Not Gemma	0.122	0.306
7	Fahras	0.227	0.522
	Baseline	*0.368*	*0.691*

Table 5: Subtask 2 corpus-level results (lower is better). All seven teams outperform the baseline.

All seven submitted systems outperformed the organiser baseline (CER 0.368, WER 0.691). Misraj AI achieved the lowest corpus-level CER of 0.079, a 78.5% relative reduction over the baseline, and the lowest WER of 0.244. The top six teams form a competitive cluster with CER between 0.079 and 0.122, separated by a gap from Fahras (0.227).

The CER and WER rankings do not always align. Oblevit ranks second by CER (0.093) but has the fourth-highest WER (0.327) among the top six, suggesting that its character-level errors are disproportionately concentrated at word boundaries. Conversely, Ketaba OCR ranks third by CER (0.094) but achieves a lower WER (0.300) than Oblevit, indicating better word-level coherence. Not Gemma (rank 6 by CER at 0.122) achieves a lower WER (0.306) than four higher-ranked teams, suggesting that its errors are distributed across characters rather than concentrated in patterns that destroy entire words.

7.3. Subtask 2: Per-Line Results

Table 6 presents the per-line CER and WER, where metrics are computed independently for each test line and averaged.

#	Team	CER	WER
1	Ketaba OCR	**0.082**	0.259
2	Misraj AI	0.090	**0.252**
3	Latent Narratives	0.100	0.285
4	Oblevit	0.105	0.332
5	Al-Warraq	0.106	0.347
6	Not Gemma	0.110	0.313
7	Fahras	0.182	0.430
	Baseline	*0.281*	*0.588*

Table 6: Subtask 2 per-line results, averaged over individual test lines (lower is better).

The per-line evaluation reveals ranking changes compared to corpus-level scoring. Most notably, Ketaba OCR rises from rank 3 (corpus) to rank 1 (per-line), achieving a per-line CER of 0.082, while Misraj AI moves from rank 1 to rank 2 (0.090). Oblevit drops from rank 2 (corpus) to rank 4 (per-line), while Latent Narratives improves from rank 4 to rank 3.

These shifts arise from the different weighting implicit in each metric. Corpus-level scoring weights each character equally, so longer lines contribute more to the aggregate. Per-line scoring weights each line equally regardless of length, giving more influence to shorter lines. The rank reversal between Misraj AI and Ketaba OCR suggests that Misraj AI performs better on longer body-text lines, while Ketaba OCR handles short and irregular lines more consistently. This distinction has practical implications: for applications requiring uniform transcription quality across an entire manuscript page—including headers, marginalia, and fragmented lines—per-line CER may be the more informative metric.

7.4. Error Analysis

We conducted a qualitative error analysis on 200 randomly sampled test lines from the top-ranked

system (Misraj AI, corpus-level). Errors were categorised into five types.

Character confusion (38% of errors). The most frequent error type involves visually similar characters, particularly confusions between *dal*/*dhal* (د/ذ), *sin*/*shin* (س/ش), and *ha*/*ta marbuta* (ه/ة). These confusions arise from the author's handwriting style, in which distinguishing dots are occasionally faint or absent.

Word boundary errors (24%). Incorrect segmentation of connected words or splitting of single words accounted for approximately one quarter of errors. This pattern is particularly common in passages where the author writes with minimal inter-word spacing.

Diacritics and dots (18%). Missing or misplaced diacritical dots affect the identity of several Arabic letters. Systems frequently omit dots that are faintly written or shifted from their canonical position relative to the baseline character.

Damaged or degraded regions (12%). Ink bleeding, page fold marks, and photocopy artefacts cause localised failures. These errors are concentrated in the earlier documents (1951–1953), where physical preservation is poorest.

Rare vocabulary and proper nouns (8%). Names of people, places, and uncommon political terminology that appear infrequently in the training data are more likely to be misrecognised, as language model priors cannot compensate for visual ambiguity in these cases.

7.5. Discussion

Baseline versus participants. Although the organiser baseline and several participating teams use the same model family (Qwen-VL variants), all seven teams achieved lower CER than the baseline. This indicates that differences in data augmentation, training strategy, prompt design, hyperparameter tuning, and ensemble methods yield substantial improvements even when the underlying architecture is shared. The gap between the baseline (CER 0.368) and the best system (CER 0.079) represents a 78.5% relative CER reduction.

Corpus-level versus per-line evaluation. The ranking shifts between Tables 5 and 6 underscore the importance of reporting both metrics. Corpus-level CER captures aggregate transcription fidelity and is appropriate for applications such as full-text search, while per-line CER better reflects consistency across heterogeneous line types. Future shared tasks should report both metrics, as they capture complementary aspects of system quality.

8. Conclusion and Future Work

The NAKBA NLP 2026 shared task engaged participants across two complementary tracks, advancing both the available ground truth and the state of automatic transcription for Arabic manuscripts. Beyond the performance results, the shared task makes several contributions to the field: the public release of a large-scale, line-level annotated Arabic manuscript corpus; a standardised evaluation framework combining manual transcription assessment with automatic system benchmarking; and baseline results that can serve as reference points for future research.

On the systems side, all seven participating teams outperformed the organiser baseline, with the top-performing system (Misraj AI) reducing corpus-level CER from 0.368 to 0.079. Rankings shift between corpus-level and per-line evaluation—Ketaba OCR achieves the lowest per-line CER (0.082) while Misraj AI leads on corpus-level CER—revealing that different systems have different strengths across line types. On the transcription side, manual annotation by trained annotators achieved CER between 0.061 and 0.114, confirming that Arabic manuscript transcription remains challenging even for human annotators. The divergence between CER and WER rankings across teams reveals qualitative differences in error patterns with implications for downstream applications.

All contributed transcriptions and system outputs are released under CC-BY-4.0 via the shared task page.[3] Future iterations will expand the dataset to additional authors and genres, introduce page-level and paragraph-level evaluation, explore cross-author generalisation, and investigate strategies for increasing participation in the manual transcription track.

Acknowledgements

We thank all participating teams for their contributions and the reviewers for their constructive feedback. We gratefully acknowledge the Arab Center for Research and Policy Studies for providing the

[3] https://acr.ps/1L9BaeY

Omar Al-Saleh Memoir Collection, and the annotators whose careful work produced the ground-truth transcriptions used in this shared task.

Ethical Considerations

The Omar Al-Saleh Memoir Collection is used with the permission of the Arab Center for Research and Policy Studies. The memoirs contain personal and political reflections from a specific historical period; participating teams were advised to handle the content with appropriate scholarly care. All teams agreed to ethical use guidelines, and no generative AI was permitted for manual transcription in Subtask 1.

Limitations

The dataset represents a single author's handwriting, which limits conclusions about generalisability to other Arabic manuscripts or scripts. Subtask 1 attracted only three participants, limiting the statistical conclusions that can be drawn about manual transcription quality. Because each Subtask 1 team transcribed a different batch, formal inter-annotator agreement metrics cannot be computed across teams. Evaluation is restricted to line-level CER and WER; document-level or page-level metrics may better capture transcription quality for downstream use.

References

O. S. Al-Barghouti and K. Totah. 1923. *Tarikh Filastin [History of Palestine]*. Bayt al-Maqdis Press, Jerusalem.

N. Alrobah and S. Alzahrani. 2022. Arabic handwritten text extraction: A survey. *IEEE Access*, 10:50437–50465.

Arab Center for Research and Policy Studies. 2024. Palestine Memory Project.

J. Bai, S. Bai, S. Yang, et al. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Z. Chen et al. 2024. InternVL: Scaling Up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*.

C. Cieri, M. Maamouri, S. Strassel, et al. 2016. MADCAT: Multilingual automatic document classification, analysis, and translation corpus.

C. Clausner, A. Antonacopoulos, N. McGregor, and D. Wilson-Nunn. 2018. ICFHR 2018 competition on recognition of historical Arabic scientific manuscripts – RASM2018. In *Proc. ICFHR*, pages 471–476.

A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of ICML*, pages 369–376.

Institute for Palestine Studies. a. Omar al-saleh al-barghouti (1894–1965). Accessed: 2025.

Institute for Palestine Studies. b. Palestine Social History Archives: Omar Saleh Al-Barghouti Diaries. Accessed: 2025.

Mustafa Jarrar, Mo El-Haj, Amal Haddad, Serin Atiani, Shadi Abudalfa, Khalil Sima'an, Paul Rayson, and Camille Mansour, editors. 2026. *Proceedings of the second International Workshop on Nakba Narratives as Language Resources*. Association for Computational Linguistics, Spain.

G. Jocher, A. Chaurasia, and J. Qiu. 2024. Ultralytics YOLO. Version 8.3.0+.

M. Kassis, A. Abdalhaleem, A. Droby, R. Alaasam, and J. El-Sana. 2017. VML-HD: The historical document dataset. In *Proc. ICDAR Workshops*.

R. Khanam and M. Hussain. 2024. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv preprint arXiv:2410.17725*.

M. Li, T. Lv, J. Chen, et al. 2023. TrOCR: Transformer-based optical character recognition with pre-trained models.

S. A. Mahmoud, I. Ahmad, W. G. Al-Khatib, et al. 2014. KHATT: An open Arabic offline handwritten text database. *Pattern Recognition*, 47(3):1004–1011.

G. Muehlberger, L. Seaward, M. Terras, et al. 2019. Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation*, 75(5):954–976.

M. Pechwitz, S. S. Maddouri, V. Märgner, et al. 2002. IFN/ENIT – database of handwritten Arabic words. In *Proc. CIFED*, pages 129–136.

M. Romanov et al. 2023. OpenITI: Open Islamicate Texts Initiative.

V. Romero, J. A. Sánchez, A. H. Toselli, et al. 2012. The ESPOSALLES database: An ancient marriage license corpus for off-line handwriting

recognition. *Pattern Recognition*, 46(6):1658–1669.

A. Saeed, C. Ascherl, N. Uchigasaki, et al. 2024. Muharaf: Manuscripts of handwritten Arabic dataset for cursive text recognition. In *NeurIPS Datasets and Benchmarks Track*.

R. Xu, Y. Wang, S. Zhang, et al. 2024. Florence-2: Advancing a unified representation for a variety of vision tasks. *arXiv preprint arXiv:2311.06242*.