QaLLM: An LLM-based NER Dataset Curation, Annotation and Evaluation in Historical Urdu Elegies

Saniya Irfan[1], and Syed Juned Ali[2]

[1] Humanities and Social Sciences, IIT Delhi, India
[2] Business Informatics Group, TU Wien, Austria

Abstract

Digital humanities increasingly use computational tools to analyze large literary corpora, yet low-resource, right-to-left languages like Urdu remain underserved, thereby, hindering research on culturally rich genres such as Marsiya, South Asia's elegiac poetry tradition. Named-entity recognition (NER) in Marsiya involves specific challenges that existing methods either ignore the genre or depend on costly manual annotation, limiting digital humanities research in the Global South. To address this, we present QaLLM, an end-to-end framework leveraging large language models (LLMs) for Urdu Marsiya NER with a human-in-the-loop validation stage. We conduct empirical analysis comparing multiple state-of-the-art LLMs and prompting configurations, and employ an LLM-as-a-Judge strategy using independent models to evaluate tagging quality. Results show that LLMs can serve as reliable first-pass annotators and reviewers, enabling efficient tagging and validation. Our contributions include — (*i*) the first publicly available Urdu Marsiya NER dataset, (*ii*) an open, reproducible methodology for low-resource, right-to-left NER with human and LLM-based validation, and (*iii*) an extensive comparative evaluation of LLMs and prompting strategies. The framework generalizes to other low-resource, complex-script languages, supporting reproducible digital scholarship and inclusive computational analysis of global literary heritage.

Keywords: Named Entity Recognition, Elegies, Large Language Model, Dataset Annotation, Digital Humanities

1 Introduction

Digital humanities scholarship has increasingly used advanced computational text-analysis methods to extract patterns and meanings in large corpora that are prohibitively labour-intensive and expensive to examine by hand. In computational text analysis, named-entity recognition (NER) is a crucial task in semantic text analysis and semantic parsing [12] that allows identifying and classifying mentions of people, places, organizations, and other semantically salient entities. NER enables scholars to reconstruct networks of influence, chart geographies of discourse, and surface intertextual references across vast collections. Reliable NER pipelines have created interpretive and data-driven studies in literary history, cultural geography, and intellectual networks [8; 24]. However, most of these successes have centered on high-resource European languages, specifically modern English, which benefits from abundant annotated corpora, off-the-shelf tools, and community standards. As digital humanists seek to broaden the scope of inquiry and extend computation methods to non-European literatures, low-resource languages pose a key challenge: the

Saniya Irfan, and Syed Juned Ali. "QaLLM: An LLM-based NER Dataset Curation, Annotation and Evaluation in Historical Urdu Elegies." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 887–902. https://doi.org/10.63744/nxwIBAfXngn3.

lack of labeled datasets to train computational tools. Limited data undermines tool performance, while script-specific challenges, such as bidirectionality and context-sensitive tokenization, increase the difficulty of adapting existing pipelines to low-resource languages with their syntactic and semantic rules.

Urdu Marsiya is elegiac poetry commemorating the tragedy of Karbala that holds a central place in South Asian Shia devotional literature; with its rhetorical devices such as repetition and contrast, archaic lexical choices, and dense allusions to religious and historical narratives, it functions both as a living ritual medium and a rich archive of cultural memory. However, its unique blend of Persian and Arabic borrowings, specialized religious terminology, and right-to-left script makes it difficult for standard NER models-trained on modern newswire or everyday prose-to correctly recognize and classify its names and terms. Moreover, despite Marsiya's importance for scholars of South Asian literature, religious studies, and performance anthropology, no publicly available, machine-readable NER corpus exists for this genre. As a result, digital-text-analytics projects either ignore Marsiya texts altogether or rely solely on human annotation-an approach that, while accurate, is prohibitively time-consuming and difficult to scale beyond small samples. Urdu remains in a family of low-resource languages that lack the availability of machine-readable texts and gold-standard, annotated datasets for core NLP tasks (e.g., NER, part-of-speech tagging). Right-to-left scripts introduce further technical barriers, such as many tokenization libraries assume left-to-right directionality, leading to misaligned character encoding, improper word boundaries, and degraded model performance. Even where Urdu NER corpora exist, existing works have primarily focused on newswire or social-media genres [2; 3], leaving domain-specific entities-such as sacred sites (e.g., "Imambara," "Dargah"), historical personages (e.g., "Imam Husayn," "Zainab bint Ali"), and ritual terms (e.g., "Majlis," "Noha")-poorly represented or absent. Off-the-shelf models thus struggle to recognize and correctly classify the full range of entities that matter for Marsiya studies. Further, the poetic and often archaizing registers of classical Marsiya employ morphological inflections and honorifics that do not appear in contemporary corpora, increasing the OOV (out-of-vocabulary) issues.

Though capable of high inter-annotator agreement, traditional manual annotation workflows require extensive scholar time and domain expertise, a bottleneck for projects aiming to assemble large-scale corpora. Conversely, fully automated pipelines promise rapid processing but risk introducing systematic biases and errors that, unchecked, could lead scholars astray in quantitative analyses. Neither extreme, fully manual nor fully automated, adequately balances the crucial imperatives of scale and quality, especially for a genre as nuanced and culturally significant as Urdu Marsiya. Therefore, our paper proposes a hybrid approach that leverages the linguistic knowledge embedded in large language models (LLMs) while retaining expert human oversight through an efficient validation interface. By automatically pre-tagging entity spans, confidence scores, and justifications for the tagged entity with an LLM and then directing human annotators' attention to low-confidence or novel cases, we aim to reduce manual effort without sacrificing annotation quality. This human-in-the-loop paradigm aims to achieve faster corpus expansion and the consistency required for rigorous digital-humanities research. Beyond the immediate benefits for Marsiya scholarship, our project is a blueprint for extending DH methodologies into other low-resource, right-to-left literary domains, such as classical Persian poetry, Ottoman archival records, or Arabic philosophical texts. By documenting best practices for dataset curation, prompt engineering, and collaborative evaluation, we seek to equip scholars working on under-studied traditions with the tools and workflows necessary to bring their corpora into computational analysis pipelines.

Concretely, in our work, we curated a representative corpus of Urdu Marsiya drawn from the canonical 19th-century poet Mir Anees. Annotation guidelines define a schema of six primary entity types: Person, Location, Organization, Date, Time, and Designation. Our automated pipeline uses existing state-of-the-art LLMs from OpenAI,[1] DeepSeek,[2] and Anthropic Claude.[3] We apply several prompting strategies from general language-agnostic NER prompting to strategies tailored to Urdu script and Marsiya's genre conventions. Prompt templates provide a brief genre context and explicit instructions for span extraction, followed by in-prompt examples from the annotated corpus. Output parsing routines convert the LLM's textual response into BIO-tag sequences compatible with standard NER evaluation tools [17]. We design a lightweight web interface that visualizes model predictions alongside NER tag confidence scores. Annotators can quickly accept high-confidence tags, edit or delete uncertain extractions, and add missing entities. Each correction is logged to provide feedback for iterative prompt refinement and potential model fine-tuning. Finally, to provide an immediate validation to the human annotator, we have incorporated the LLM-as-a-judge method where multiple LLMs independently critique model outputs by evaluating predicted entities based on a well-defined criteria, offering categorical judgments (Correct, Partial, Incorrect) and rationale. Aggregating these judgments yields an alternate performance profile that we assess for alignment with human evaluation, exploring both the promise and limitations of automated judgment.

In order to evaluate our framework, we focus on the following research questions —

[**RQ1**] How accurately can a state-of-the-art LLM perform NER on Urdu Marsiya relative to expert human annotation?

[**RQ2**] How do different prompting strategies affect the LLMs performance?

[**RQ3**] How does the LLM-as-a-judge approach compare to human validation and judgment?

The remainder of this paper is structured as follows: In Section 2, we elaborate on the literature related to our work. Section 3 presents our end-to-end NER framework and all its components. We evaluate the performance of our framework in Section 4. We discuss our results and threats to the validity of our work in Section. 5. We conclude our paper with future work in Section. 6. All code covering data ingestion, prompt generation, annotation interface, and evaluation scripts is released under a permissive open-source license.[4]

2 Related Work

In the following, we provide relevant literature to our work and the background that forms the basis of our pipeline and subsequent evaluation of our framework. First, we discuss NER in DH in general and for low-resource languages. Next, we discuss domain-specific NER. After that, we discuss the human-in-the-loop annotation works. Then we discuss using pretrained language models in the literature for NER. Finally, we elaborate a bit on LLM-as-a-Judge evaluation strategies used in the literature.

[1] `https://chat.openai.com/`
[2] `https://chat.deepseek.com/`
[3] `https://claude.ai/`
[4] `https://github.com/junaidiiith/urdu-marsiya-ner-annotator`

2.1 Named-Entity Recognition in Digital Humanities

Large-scale NER has become a fundamental tool in digital humanities for uncovering social, geographic, and intertextual linkages within literary and historical corpora. Ehrmann et al. (2021) survey the challenges of applying NER to historical documents, evolving orthography, and domain-specific vocabulary, and review existing resources and approaches in the humanities context [7]. More recently, Weijers and Bloem (2025) evaluate several mainstream NER systems (e.g., spaCy, [5] Stanford CoreNLP [6]) on 20th-century philosophical texts, demonstrating that off-the-shelf models frequently miss or misclassify specialized entity types, thereby underscoring the need for custom annotation and modeling workflows in DH projects [23].

2.2 NER for Low-Resource and Right-to-Left Languages

Urdu NER has progressed from feature-based statistical models to neural architectures. Kanwal et al. [11] introduced the MK-PUCIT corpus, experimenting with Word2Vec [15], fastText [10], and RNNs. Subsequent work integrated subword embeddings (Floret) with BiLSTM, GRU, and CRF models [4], while attention-based BiLSTM-CRF approaches improved entity focus [21]. UNER-II further expanded corpus coverage using contextual augmentation [20]. In Arabic, the Wojood corpus [9] achieved high-quality nested annotation (micro-$F_1 = 0.884$) using AraBERT. For Shahmukhi-script Punjabi, contextual embeddings and transfer learning improved performance [19], and multimodal Urdu NER benchmarks (U-MNER) aligned visual and textual cues for right-to-left languages [1].

2.3 Domain-Specific NER in Literary Studies

Domain adaptation studies highlight that genre and period profoundly influence NER performance. Weijers and Bloem's (2025) work on philosophical texts reveals frequent misclassifications of specialized person and concept names-issues that likely parallel the challenges of archaic, allusive vocabulary in Marsiya poetry [23]). Although NER for poetry remains under-explored, studies of domain-specific NER in other literary genres (e.g., classical Persian) suggest that tailored annotation schemas and models are essential for capturing genre-specific entities.

2.4 Human-in-the-Loop Annotation

Interactive annotation tools combine automation with expert oversight to accelerate corpus creation. Brat [18] supports text-bound annotations and has inspired enhanced systems such as Markup [6], which employs active learning, and SciAnnotate [14], which integrates weak-label sources for efficiency gains. NeuroNER [5] links annotation, model training, and prediction, illustrating how human-in-the-loop design sustains both speed and quality in NER workflows.

2.5 Pretrained and LLM-based NER

Transformer architectures [22] revolutionized low-resource NER through cross-lingual transfer and few-shot prompting. Surveys such as Keraghel et al. [12] highlight how reinforcement-learning fine-tuning and graph neural networks enhance label consistency across heterogeneous corpora. Collectively, these studies demonstrate that pretrained LLMs can generalize well to new domains when appropriately prompted or adapted.

[5] `http://spacy.io/`
[6] `https://techfinder.stanford.edu/technology/stanford-corenlp`

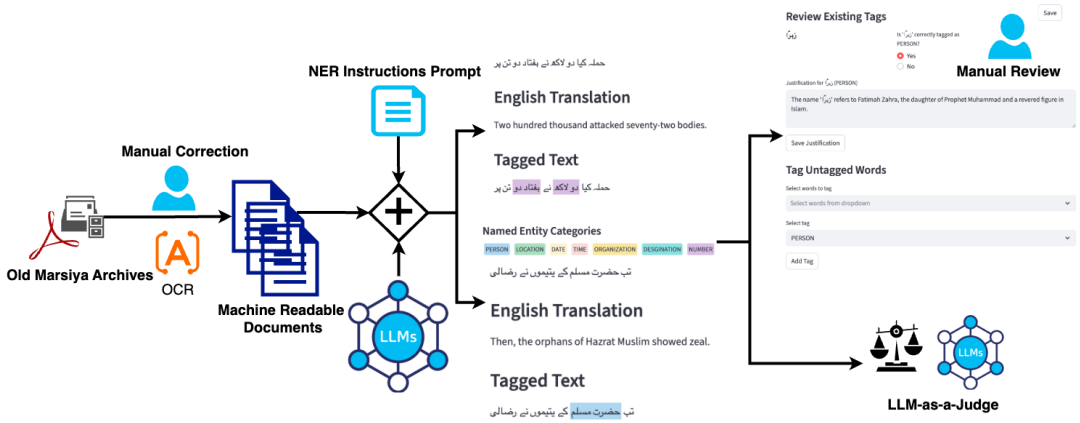


Figure 1: QaLLM Framework: A High Level Overview

2.6 LLM-as-a-Judge Evaluation

LLM-based evaluators provide scalable alternatives to human assessment. Best-practice overviews [13; 16] and industry guidelines (Evidently AI[7]) document design strategies, bias risks, and reliability metrics. Recent studies [13; 16] show strong alignment between LLM judgments and human annotations, supporting the viability of LLM-as-a-Judge for large-scale, reproducible evaluation.

2.7 Synopsis

Based on the literature, we note a gap in the literature for supporting NER in low-resource languages; however, substantial, promising work has been done to bridge this gap. Pretrained language models like LLMs have shown promising results, and LLM-as-a-judge has also proven useful in demonstrating alignment with human annotations in open-ended benchmarks. Based on our literature review, we developed our LLM-based NER pipeline, i.e., QaLLM, and augmented the framework with an LLM-as-a-judge evaluation package.

3 QaLLM Framework

In the following, we elaborate on the end-to-end QaLLM framework as shown in the Figure. 1. **Machine Readable Corpus Generation** details the extraction of Urdu *Marsiya* texts, OCR processing, and normalization steps required to prepare a clean digital corpus. **Annotation Schema and Guidelines** defines the entity ontology, model selection process, and prompt engineering strategy used to fine-tune LLM-based tagging. **Human-in-the-Loop Annotation Interface** describes the design of the annotation platform that enables collaborative expert review and justification logging. **LLM-as-a-Judge Evaluation Framework** introduces the multi-LLM evaluation mechanism used to assess the correctness of entity tags. Finally, **Data Export and Tool Accessibility** explains how the annotated datasets and associated resources are serialized, versioned, and made publicly available for reuse and extension.

3.1 Machine Readable Corpus Generation

In the following, we elaborate on the corpus generation from a dataset of Urdu elegies.

[7] `https://www.evidentlyai.com/llm-guide/llm-as-a-judge`

OCR Text Extraction. We began by extracting 242 Urdu Marsiya texts by the nineteenth-century poet Mir Anees from the eMarsiya online library[8]. The present corpus focuses on Mir Anees, whose canonical Marsiya corpus offers rich linguistic diversity and stable authorship. This controlled scope facilitates consistent evaluation; however, future iterations will extend the corpus to include contemporaneous poets such as Mirza Dabeer and later revivalists to assess genre-level generalization.

Since these texts exist only as scanned images, we applied Google's Vision API[9] for optical character recognition (OCR) to produce machine-readable UTF-8 text files.

Extraction, Manual Inspection and Preprocessing. Many scanned images did not have horizontally aligned text, but instead diagonally written text, which caused OCR to perform poorly. Moreover, given the low out-of-the-box accuracy of Urdu OCR, we involved manual inspection and correction by expert annotators over two months. This labor-intensive step ensured high-quality input for downstream processing and highlighted the urgent need for improved OCR techniques for low-resource, right-to-left scripts.

Preprocessing and Normalization. The corrected text then passed through a multi-stage cleaning pipeline as follows: *i*) **Unicode Normalization (NFC):** Standardize Nastaliq script encoding to avoid spurious character variants. *ii*) **Diacritic and Whitespace Cleanup:** Remove extraneous Arabic–Persian diacritics and fix OCR-induced spaces or line breaks. *iii*) **Segmentation:** Split text into verses and lines using rule-based heuristics derived from Marsiya's metrical patterns. *iv*) **Punctuation Normalization:** Align Arabic–Persian punctuation marks with Urdu typographic norms and *v*) **Sentence Boundary Detection:** Combine rule-based token patterns with a small manually annotated corpus to accurately delineate verse-level units for annotation.

3.2 Annotation Schema and Guidelines

Once we extracted and preprocessed the data, we needed to do human annotation that would act as ground truth for evaluation. We defined seven primary entity classes that we aimed to extract as named entities from the Urdu text. Table 1 formalizes the ontology with definitions, examples, and edge-case guidance.

Model Selection and Prompt Engineering. We selected four closed-source and two open-source LLMs. In case of closed source LLMs for LLM-based NER tagging, we used GPT4o, GPT4o-mini, GPT4.1, GPT4.1-mini, and in case of open-sourced, we selected both the most advanced versions of DeepSeek, namely, deepseek-chat and deepseek-reasoner. We selected different LLMs from the same provider to evaluate the effect of different LLMs from the same provider. Next, in order to further fine-tune LLM-based NER, we crafted four different prompt types in increasing order of Urdu Marsiya-specific context detail and complexity—

1. **General Urdu NER Prompt:** "Tag all named entities in this Urdu text."
2. **Genre Context Prompt:** "Marsiya poetry often includes Karbala references-tag names, places, and dates accordingly."
3. **Urdu-Script Context Prompt:**
    - Remind the model that Urdu runs right-to-left.
    - Recognize Urdu–Persian loanwords (e.g., یزید, شعیب (*Shoaib, Yazid*)) as PERSON when they are names.

[8] https://emarsiya.com/

[9] https://cloud.google.com/vision/docs/ocr

Table 1: Entity Ontology used in QaLLM with definitions, examples, and labeling notes. Urdu examples include English glosses.

Entity	Definition	Urdu Example	Notes
Person	Named individual or mytho-religious figure.	علی ابن حسین (Husayn ibn ᶜAlī), حضرت عباس (Hazrat ᶜAbbās)	Honorifics may attach (حضرت, سید); if honorific stands *without* a name and denotes rank, prefer DESIGNATION.
Location	Geographical or sacred place; shrine/maqām; battle site.	کربلا (Karbala), حسین امام حرم (Sanctuary of Imam Husayn)	Shrine compounds treated as LOCATION; if an institution (e.g., seminary), see ORGANIZATION.
Organization	Named institution, group, or body (religious, educational, political).	دینیہ مدرسہ (Religious seminary), عباسیہ انجمن (Abbāsīya Anjuman)	Distinguish from sacred places (LOCATION); prefer ORG when collective/institutional identity is primary.
Date	Calendar dates, months, commemorative days.	محرم (Muḥarram), محرم دس (10th of Muḥarram)	Lunar months and named observances are DATE.
Time	Times of day/periods/durations.	صیام پہر (period of fasting), سحر (dawn)	Archaic or poetic temporal expressions map to TIME.
Designation	Honorifics, titles, or epithets denoting rank/status rather than identity.	شہداء سید (Master of Martyrs), امیرالمؤمنین (Commander of the Faithful)	If attached to a specific name (امام حسینؑ) and functioning as part of the name, favor PERSON; standalone epithets are DESIGNATION.
Number	Cardinals/ordinals used as quantities.	چالیس (forty), تیسرا (third)	Numeric tokens that are part of dates should be DATE; otherwise NUMBER.

- Correctly handle zero-width non-joiners used in Urdu compound words (e.g., خوداعتمادی *self-confidence*) to prevent erroneous character joining during tokenization and tagging.

4. **Marsiya-Specific Context Prompt:**

- Identify Battle of Karbala participants (e.g., زینب عباس, (*Abbas, Zaynab*)) as PERSON or LOCATION.
- Tag shrine compounds (e.g., حسین امام حرم (*Sanctuary of Imam Husayn*)) as LOCATION.
- Recognize poetic epithets (e.g., شہداء سید (*Master of Martyrs*)) as PERSON when attached to a name, else as DESIGNATION.
- Tag archaic time expressions (e.g., صیام پہر (*period/hour of fasting*)) as TIME.
- Treat religious month references (e.g., صفر محرم, (*Muharram, Safar*)) as DATE.

Each prompt includes a single in-prompt example and requests output in our JSON schema:

```
[
  {
    "original": "<Urdu line>",
    "tagged" : "<line with <LABEL>…</LABEL> spans>",
    "translation": "<English translation>"
  },
  …
]
```

We initially tested paragraph-level inputs (5-10 lines) to provide broader context, but LLMs frequently skipped or truncated lines due to token-limit and attention-decay effects,

especially in right-to-left Urdu text. This led to missing entities and unstable span offsets. Line-level segmentation produced more consistent BIO alignment and reproducible outputs, so we adopt it as the standard granularity for QaLLM. The line-level segmentation does use a configurable number of contextual lines to provide relevant context to the LLM for NER.

Output Parsing and Tag Conversion. Once we have the LLM output, the LLM's JSON response is parsed to extract bracketed spans and map them to token-level BIO tags aligned with the original text. In the BIO tagging scheme for named-entity recognition (NER), each token in a sentence is labeled as either **B** (Beginning), **I** (Inside), or **O** (Outside) of an entity. A **B-TAG** label marks the first word of a multi-word entity (e.g., "B-PERSON" for the first word of a person's name), while **I-TAG** labels mark each subsequent word inside that same entity (e.g., "I-PERSON" for the remaining words in the name). Tokens that do not belong to any entity receive the **O** label. For example, in the sentence "Mir Anees recited Marsiya," the tokens would be tagged as "Mir B-PERSON," "Anees I-PERSON," "recited O," "Marsiya B-WORK_OF_ART." This convention ensures that entity boundaries and types are unambiguously encoded, making it straightforward to train and evaluate NER models on sequence-labeling tasks. Using the LLM output's "tagged" field, we can generate the dataset in standardized BIO-format. We make the manually curated dataset in BIO-format for researchers to use publicly available[10].

3.3 Human-in-the-Loop Annotation Interface

We implemented a Streamlit[11]-based front end to make our annotation tool accessible to users. The interface retrieves model-generated tags, displays them inline with the Urdu Nastaliq script and in English, and saves the reviewed entities. Annotators navigate the text one line at a time, with entities highlighted in distinct colors per class. For each highlighted span, users can *accept*, *reject*, *modify* its class, or *add* a missing tag. A keyboard-driven review mode accelerates common actions, while tooltips provide definitions and examples for each entity type. Furthermore, a justification for each entity is provided by default from the LLM that can support the reviewer in accepting or rejecting the NER tag. The user can update the justification, too, in case the reviewer disagrees with the justification. Our tool supports authentication; therefore, the work of one reviewer is stored separately. Logging reviewer justification enables a rich log of reviewer decisions and rationale that can later be used to analyze inter-reviewer agreement or issues. Changes are committed in real time.

3.4 LLM-as-a-Judge Evaluation Framework

To evaluate NER outputs, we deploy multiple LLMs as judges. We use Anthropic's Claude 3.7, OpenAI's GPT-4o, and GPT-4.1-mini for LLM-as-a-judge. LLMs are given an instruction prompt to judge an LLM-based NER-tagged sentence and are asked to judge the correctness of the tag. Each judge classifies each entity as *correct* or *incorrect*, and provides a rationale. In case of incorrectness, LLM provides an alternative entity. To make the judgment strict, all the judges must agree on the tagged sentence so that the NER tags can be judged correctly. However, this is a configurable parameter. The user can put a threshold indicating, out of all the LLMs, how many LLMs need to agree for the tags in the original sentence to be considered correct.

[10] `https://huggingface.co/datasets/junaidiiith/marsiya-ner-dataset`
[11] `https://streamlit.io/`

Table 2: Comparison of entity counts between the human-annotated test set and the corpus tagged by GPT-4.1-mini

Entity Type	Total Human	Unique Human	Total LLM	Unique LLM
NUMBER	237	125	2242	595
PERSON	2128	778	30321	5588
LOCATION	378	173	3307	830
TIME	226	118	1615	584
DESIGNATION	353	210	3705	1299
ORGANIZATION	50	40	453	208
DATE	78	51	596	249
Total	**3450**	**1495**	**42239**	**9353**

3.5 Data Export and Tool Accessibility

Annotated outputs are serialized in a structured JSON schema capturing token indices, entity labels, confidence scores, and annotator decisions. For interoperability, we provide an export utility that transforms JSON into Excel spreadsheets with separate sheets for each document and aggregated metadata.

4 Evaluation

In this section, we present two complementary experiments designed to evaluate the performance and scalability of the **QaLLM** framework: (1) a direct comparison of multiple LLMs against a gold-standard, human-annotated Urdu *Marsiya* NER dataset, and (2) an evaluation of a large, automatically tagged corpus using the LLM-as-a-Judge strategy to approximate annotation quality at scale.

4.1 Experimental Setup

We curated and manually reviewed a gold-standard dataset containing 3,450 entity mentions (1,495 unique) from the Mir Anees *Marsiya* corpus. To complement this, we generated a large-scale machine-annotated version of the corpus using the best-performing LLM–prompt configuration (GPT-4.1-mini with Marsiya-specific context). The resulting corpus contains 42,239 total entity tags (9,353 unique). Table 2 summarizes entity distributions across both datasets, allowing a direct comparison between human and model-generated annotations.

To assess annotation quality at scale, the LLM-tagged corpus is evaluated using the **LLM-as-a-Judge** framework (Experiment 2), where multiple models independently assess the correctness of each entity tag.

For quantitative evaluation, we compute metrics under two matching conditions:

- **Exact match:** both the text span and the entity label must coincide.
- **Fuzzy match:** the entity label must match even if the text span differs.

Precision, Recall, and F_1 scores are calculated for both conditions to provide a comprehensive measure of NER performance.

4.2 Results

Experiment 1: LLM-based NER Performance We show the overall performance of the different LLMs using their best-performing prompt in Table 3, that shows fuzzy-match Precision, Recall, and F_1 for each model. We see that GPT-4.1-mini achieves the highest F_1 (0.75), balancing strong Precision (0.74) and Recall (0.76). GPT-4.1 attains high Recall (0.71) but suffers in Precision (0.48), indicating a tendency to over-tag. Deepseek variants trade off Precision and Recall more evenly, with deepseek-chat favoring Recall and

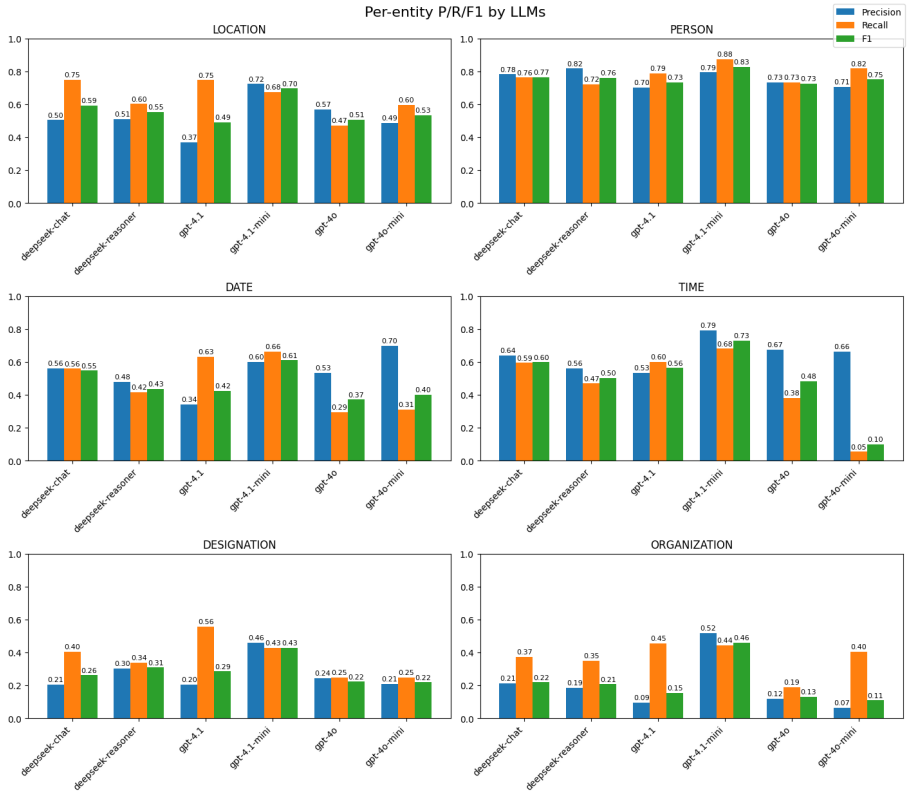


Figure 2: Per-Entity QaLLM Performance against manually annotated dataset

deepseek-reasoner favoring Precision. Figure 2 breaks down F_1 by entity type. All models excel on PERSON (max F_1 =0.83, GPT-4.1-mini) and struggle most on ORGANIZATION (F_1 $\leq$0.46) and DESIGNATION (F_1 $\leq$0.43). PERSON and LOCATION are well learned across LLMs, likely due to their prominence in training data and genre context. In contrast, rarer or domain-specific labels (ORGANIZATION, DESIGNATION) remain challenging, underscoring the need for targeted schema examples or fine-tuning on Marsiya-specific glossaries.

LLM	Precision	Recall	F_1
deepseek-chat	0.60	0.68	0.63
deepseek-reasoner	0.67	0.62	0.64
GPT-4o	0.60	0.58	0.59
GPT-4o-mini	0.59	0.63	0.60
GPT-4.1	0.48	0.71	0.57
GPT-4.1-mini	**0.74**	**0.76**	**0.75**

Table 3: Fuzzy-match overall Precision, Recall, and F_1 against human annotations.

We show the trend of LLM performance with different prompt types and LLMs in Fig. 3a and Fig. 3b. The figure shows that for both exact- and fuzzy-match evaluations, adding Marsiya-specific context to prompts consistently boosts all LLMs' recall-and often their precision-relative to more generic prompts, and the "mini" versions of GPT-4.1 outperform their full–size counterparts in balanced F_1 performance. In contrast, the GPT-4o family shows more mixed trends.

Moving from a generic "Tag all named entities" prompt to one mentioning Karbala and Marsiya conventions causes a small precision drop (2–5 points) but a larger recall drop (5–8 points exact, 4–6 fuzzy), as models adopt a more conservative stance when reminded

of genre nuances. By contrast, upgrading from a Marsiya-context to a Marsiya-specific prompt-explicitly citing shrine names, poetic epithets, and archaic time expressions-yields dramatic improvements (e.g., GPT-4.1-mini gains +12 precision and +14 recall exact, boosting F_1 from 0.54 to 0.68; deepseek-chat/reasoner see +10 recall), showing that even black-box models benefit from detailed domain cues. Finally, appending Urdu-script tips (zero-width joiners, RTL reminders) delivers only marginal or negative returns (precision –2–4points, recall flat), suggesting script-level hints can introduce span-boundary ambiguities without uncovering many new entities. In sum, carefully crafted, genre-aware prompts are the most effective way to improve LLM-based NER on Urdu Marsiya. At the same time, low-level script instructions add little and may conflict with a model's internal tokenization.

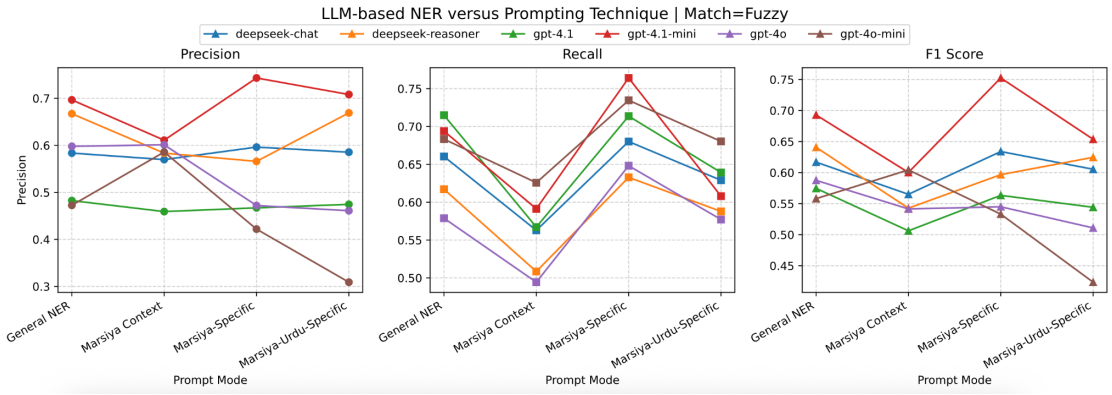
GPT-4.1-mini consistently outperforms its full-size model by a wide margin in precision-up to 20 points higher across prompts-while matching or slightly exceeding GPT-4.1's recall under Marsiya-specific prompts. This suggests that the distilled "mini" architecture retains robust entity-extraction abilities but is less prone to over-generation, making it especially valuable when annotation quality is crucial. In contrast, the GPT-4o family (both standard and mini) delivers only moderate results, with precision in the 40–55 percent range and recall between 45–60 percent under exact matching; their fuzzy-match F_1 tops out around 0.63 even with the best prompts, indicating limitations in handling right-to-left scripts and domain-specific cues. The deepseek models occupy a middle ground: deepseek-chat favors recall (approaching 0.69 exact under Marsiya-specific prompts), whereas deepseek-reasoner leans toward precision (around 0.67), yielding balanced F_1 scores of 0.61–0.64 exact (0.67–0.69 fuzzy). Overall, GPT-4.1-mini strikes the strongest precision–recall trade-off for automated pre-annotation. At the same time, GPT-4o's underperformance underscores that latest LLMs are not automatically best suited for low-resource, right-to-left genres without targeted adaptation.

Across every model and prompt variant, fuzzy-match F_1 scores exceed exact-match by roughly 5–10 points. This highlights that span-boundary disagreements are a common challenge when tagging dense, poetic text but often benign when labels themselves are correct. Crucially, the relative ranking of models remains stable between matching modes: GPT-4.1-mini leads, deepseek models form a middle tier, and GPT-4o variants lag. This consistency suggests that researchers can reliably use fuzzy-match metrics for comparative evaluation, confident that they mirror exact-match trends while forgiving minor boundary offsets. Nonetheless, final corpus validation should still include exact-match checks on a human-reviewed subset to ensure that boundary precision meets rigorous quality standards.

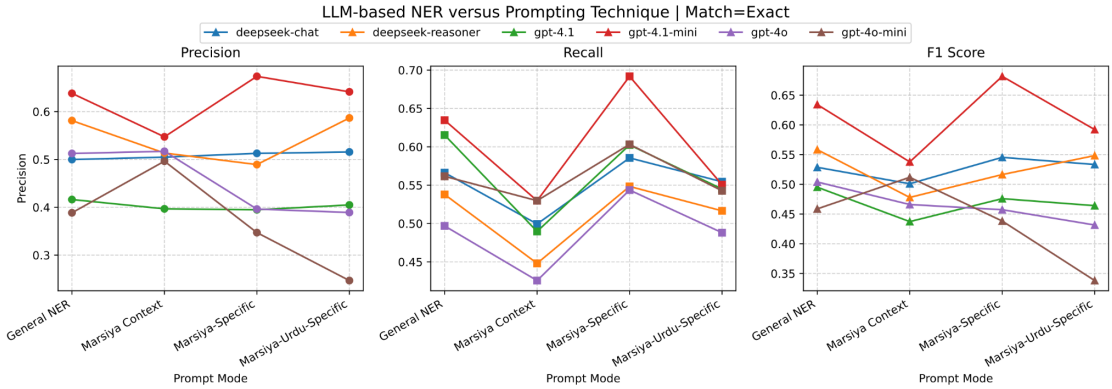
Experiment 2: LLM-as-a-Judge on the Full QaLLM Corpus

To scale our evaluation beyond the hand-annotated subset, we treated the GPT-4.1-mini–tagged corpus (42,239 entity mentions) as a pseudo-gold standard and enlisted three judge LLMs—Anthropic Claude 3.7 Sonnet, GPT-4o-mini, and GPT-4.1—to re-label every span. We then computed fuzzy-match F_1 (agreement) between each judge and the original GPT-4.1-mini annotations.

The combined inter-judge agreement was 64.37%, meaning that all three LLMs concurred on roughly two-thirds of the annotations. Interestingly, this mirrors the model's average F_1 ($\approx$65%) against the human-annotated gold set (see Table 3). The convergence of these two results, i.e., human vs. model and model vs. model—suggests that LLMs exhibit similar patterns of confidence, disagreement, and boundary error as human annotators. In other words, the "noise profile'' of LLM judgments approximates human variability, supporting their use as scalable secondary evaluators in low-resource settings.



(a) Fuzzy-match performance across prompt strategies.



(b) Exact-match performance across prompt strategies.

Figure 3: Comparison of LLM performance under different prompt configurations for Urdu *Marsiya* NER. Both fuzzy- and exact-match evaluations show that domain-specific prompts substantially improve tagging precision and recall across all models.

When asked to critique GPT-4.1-mini's annotations on the full corpus, GPT-4o-mini emerged as the most consistent judge ($F_1 = 0.78$), followed by GPT-4.1 (0.74) and Claude 3.7 (0.58). Agreement patterns by entity type further reinforce the alignment between LLM-judge and human evaluation. Both setups show high reliability for PERSON and LOCATION, moderate stability for DESIGNATION, and poor consistency for ORGANIZATION—the same trend observed in Table 3. This near-identical ranking of entity-level difficulty across human and LLM evaluations provides quantitative evidence that the LLM-as-a-Judge paradigm captures the same underlying judgment dynamics as human annotators.

Drilling down by model, GPT-4o-mini excels on LOCATION ($F_1 \approx 0.90$) and PERSON (0.81) but underperforms on TIME (0.57) and ORGANIZATION (0.53), underscoring its strong span recognition but weaker temporal sensitivity. GPT-4.1 achieves the highest alignment on DESIGNATION (0.86), echoing its superior human-evaluated performance on the same category. Claude 3.7 produces more uniform but lower-precision judgments, consistent with its more conservative tagging tendencies. Collectively, these parallels suggest that LLM judges reproduce human-like selectivity and bias—over-tagging common entities, under-detecting rare ones—thus validating their utility as cost-efficient, reproducible proxies for human adjudication in large-scale annotation pipelines.

Responses to the research questions Based on the experiments reported above, we summarize our findings as follows:

Entity	GPT-4o-mini	GPT-4.1	Claude 3.7
DATE	0.63	**0.67**	0.51
PERSON	**0.81**	0.71	0.53
LOCATION	0.90	**0.84**	0.78
TIME	0.57	**0.69**	0.52
ORGANIZATION	**0.53**	0.51	0.30
DESIGNATION	0.72	**0.86**	0.73
Overall F_1	0.78	0.74	0.58

Table 4: LLM-as-a-Judge fuzzy-match F_1 against GPT-4.1-mini annotations.

[**RQ1**] Large language models can perform NER on Urdu *Marsiya* with accuracy approaching expert human annotation, achieving up to 0.75 F_1 under fuzzy matching (GPT-4.1-mini). This demonstrates that, when properly prompted, LLMs can serve as reliable first-pass annotators even in low-resource, right-to-left literary domains.

[**RQ2**] Prompt design exerts a decisive influence on model performance. Domain-specific prompts incorporating *Marsiya*-specific context (e.g., Karbala participants, shrine compounds, archaic time expressions) significantly improved both precision and recall across all LLMs, while script-level hints alone provided marginal benefit. This highlights that contextual grounding, not merely linguistic cues, is key to effective NER in poetic and historical texts.

[**RQ3**] The LLM-as-a-Judge evaluation closely mirrors human assessment patterns, with similar overall F_1 ($\approx$65%) and identical entity-wise difficulty rankings (PERSON, LOCATION > DESIGNATION > ORGANIZATION). This indicates that LLM judges capture the similar confidence and disagreement profiles observed in human annotation, validating their use as scalable, low-cost proxies for human adjudication in large-scale corpus creation.

5 Discussion

The evaluation demonstrates that carefully prompted and domain-informed LLMs can significantly enhance Named Entity Recognition (NER) performance in specialized literary domains like Urdu Marsiya poetry. GPT-4.1-mini notably excels, balancing precision and recall effectively, underscoring the value of distilled models that avoid over-generation issues common in their larger counterparts. Furthermore, the performance gain from detailed Marsiya-specific prompts highlights the importance of incorporating deep cultural and literary context in digital humanities (DH) research. These findings reveal that although modern LLMs have impressive generalization capabilities, domain specificity-such as poetic genres, historical context, and archaic language-remains a critical factor, necessitating careful prompt engineering or targeted fine-tuning for optimal outcomes.

Applying LLM-as-a-judge to scale annotation validation exemplifies an impactful methodology for DH projects facing limitations in human annotation resources. This approach reduces human effort and introduces a systematic, reproducible method for annotation verification. However, the variance in model agreement-particularly for ambiguous entities like organizations and temporal expressions-highlights ongoing challenges in automated NER for humanities texts. Future work should focus on creating clearer schema definitions, leveraging ensemble models for enhanced accuracy, and developing domain-specific fine-tuning datasets. Ultimately, such advancements will further establish scalable, reliable computational methods integral to humanities scholarship, enabling richer analytical insights across diverse linguistic and cultural corpora. By combining these insights, QaLLM can be tuned to maximize both throughput and accuracy, providing a robust blueprint for low-resource, RTL literary NER.

Humanistic Applications. Beyond methodological support, QaLLM can enable new humanities research avenues using the named entities data like tracing how elegies invoke familial or geographic networks surrounding Karbala, mapping shifts in devotional vocabulary across poets, and studying how sacred space and lineage structure Urdu poetics. These illustrate how computational annotation can deepen literary-historical inquiry.
Limitations. While QaLLM demonstrates strong potential, several limitations remain. First, the corpus currently focuses on the works of a single poet (Mir Anees) which limits its generalizability across the broader Marsiya tradition. Second, the LLM-as-a-Judge evaluation primarily measures precision rather than full recall, and while its agreement trends mirror human judgment, a formal human–LLM correlation study on a larger subset would strengthen validation. Third, cost and API constraints restricted systematic tuning of open-source baselines, and differences in LLM tokenization for right-to-left scripts may still introduce subtle boundary errors. These constraints define directions for future refinement rather than fundamental limitations of the framework.

6 Conclusion and Future Work

In this paper, we introduced QaLLM, an end-to-end named-entity recognition framework explicitly tailored for Urdu Marsiya, addressing significant challenges posed by low-resource, right-to-left literary texts. Our hybrid approach leverages advanced large language models (LLMs) augmented by human-in-the-loop validation and LLM-as-a-judge evaluation, effectively balancing annotation quality and scalability. Through comprehensive empirical evaluations, we demonstrated that domain-informed prompts significantly enhance LLM performance, with GPT-4.1-mini emerging as particularly robust due to its optimal precision–recall balance, and employing LLMs as scalable validation judges showed promise, though also highlighted persistent ambiguities that necessitate clearer annotation guidelines and ensemble strategies.

Our work advances Urdu literary scholarship by lowering barriers to computational analysis of culturally significant texts. Future directions involve refining annotation schema specificity, developing Marsiya-specific fine-tuned models, and expanding the human–AI collaborative workflow to encompass additional low-resource literary traditions, enabling more inclusive, equitable, and rigorous digital humanities research.

References

[1] Ahmad, Hussain, Zeng, Qingyang, and Wan, Jing. "A Benchmark Dataset and a Framework for Urdu Multimodal Named Entity Recognition". In: *IEEE Access 13* (2025), pp. 100904–100919. DOI: 10.1109/ACCESS.2025.3576784.

[2] Ali, Abbas Raza and Ijaz, Maliha. "Urdu text classification". In: *Proceedings of the 7th International Conference on Frontiers of Information Technology*. FIT '09. 2009. DOI: 10.1145/1838002.1838025.

[3] Amjad, Maaz, Sidorov, Grigori, Zhila, Alisa, Gomez-Adorno, Helena, Voronkov, Ilia, and Gelbukh, Alexander. "Bend the Truth: A Benchmark Dataset for Fake News Detection in Urdu and Its Evaluation". In: *Journal of Intelligent & Fuzzy Systems 39*, no. 2 (2020), pp. 2457–2469. DOI: 10.3233/JIFS-179905.

[4] Anam, Rimsha, Anwar, Muhammad Waqas, Jamal, Muhammad Hasan, Bajwa, Usama Ijaz, Diez, Isabel de la Torre, Alvarado, Eduardo Silva, Flores, Emmanuel Soriano, and Ashraf, Imran. "A deep learning approach for Named Entity Recognition in Urdu language". In: *Plos one 19*, no. 3 (2024), e0300725. DOI: 10.1371/journal.pone.0300725.

[5] Dernoncourt, Franck, Lee, Ji Young, and Szolovits, Peter. "NeuroNER: an easy-to-use program for named-entity recognition based on neural networks". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2017, pp. 97–102. DOI: 10.18653/v1/D17-2017.

[6] Dobbie, Samuel, Strafford, Huw, Pickrell, W Owen, Fonferko-Shadrach, Beata, Jones, Carys, Akbari, Ashley, Thompson, Simon, and Lacey, Arron. "Markup: a web-based annotation tool powered by active learning". In: *Frontiers in Digital Health 3* (2021), p. 598916. DOI: 10.3389/fdgth.2021.598916.

[7] Ehrmann, Maud, Hamdi, Ahmed, Pontes, Elvys Linhares, Romanello, Matteo, and Doucet, Antoine. "Named entity recognition and classification in historical documents: A survey". In: *ACM Computing Surveys 56*, no. 2 (2023), pp. 1–47. DOI: 10.1145/3604931.

[8] Fields, Sam, Cole, Camille Lyans, Oei, Catherine, and Chen, Annie T. "Using named entity recognition and network analysis to distinguish personal networks from the social milieu in nineteenth-century Ottoman–Iraqi personal diaries". In: *Digital Scholarship in the Humanities 38*, no. 1 (2023), pp. 66–86. DOI: 10.1093/llc/fqac047.

[9] Jarrar, Mustafa, Khalilia, Mohammed, and Ghanem, Sana. "Wojood: Nested Arabic Named Entity Corpus and Recognition using BERT". In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 3626–3636. URL: https://aclanthology.org/2022.lrec-1.387/.

[10] Joulin, Armand, Grave, Edouard, Bojanowski, Piotr, Douze, Matthijs, Jégou, Hérve, and Mikolov, Tomas. "Fasttext. zip: Compressing text classification models". In: *arXiv preprint arXiv:1612.03651* (2016). URL: http://arxiv.org/abs/1612.03651.

[11] Kanwal, Safia, Malik, Kamran, Shahzad, Khurram, Aslam, Faisal, and Nawaz, Zubair. "Urdu named entity recognition: Corpus generation and deep learning applications". In: *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 19*, no. 1 (2019), pp. 1–13. DOI: 10.1145/3329710.

[12] Li, Jing, Sun, Aixin, Han, Jianglei, and Li, Chenliang. "A survey on deep learning for named entity recognition". In: *IEEE transactions on knowledge and data engineering 34*, no. 1 (2020), pp. 50–70. DOI: 10.1109/TKDE.2020.2981314.

[13] Li, Yuran, Mohamud, Jama Hussein, Sun, Chongren, Wu, Di, and Boulet, Benoit. "Leveraging llms as meta-judges: A multi-agent framework for evaluating llm judgments". In: *arXiv preprint arXiv:2504.17087* (2025). DOI: 10.48550/ARXIV.2504.17087.

[14] Liu, Mengyang, Luo, Haozheng, Thong, Leonard, Li, Yinghao, Zhang, Chao, and Song, Le. "Sciannotate: A tool for integrating weak labeling sources for sequence labeling". In: *arXiv preprint arXiv:2208.10241* (2022). DOI: 10.48550/ARXIV.2208.10241.

[15] Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543. DOI: 10.3115/V1/D14-1162.

[16] Raju, Ravi, Jain, Swayambhoo, Li, Bo, Li, Jonathan, and Thakker, Urmish. "Constructing domain-specific evaluation sets for llm-as-a-judge". In: *arXiv preprint arXiv:2408.08808* (2024). DOI: 10.48550/arXiv.2408.08808.

[17] Ramshaw, Lance A and Marcus, Mitchell P. "Text chunking using transformation-based learning". In: *Natural language processing using very large corpora*. Springer, 1999, pp. 157–176. DOI: 10.1007/978-94-017-2390-9_10.

[18] Stenetorp, Pontus, Pyysalo, Sampo, Topić, Goran, Ohta, Tomoko, Ananiadou, Sophia, and Tsujii, Jun'ichi. "BRAT: a web-based tool for NLP-assisted text annotation". In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 2012, pp. 102–107. URL: https://aclanthology.org/E12-2021/.

[19] Tehseen, Amina, Ehsan, Toqeer, Liaqat, Hannan Bin, Kong, Xiangjie, Ali, Amjad, and Al-Fuqaha, Ala. "Shahmukhi named entity recognition by using contextualized word embeddings". In: *Expert Systems with Applications 229* (2023), p. 120489. DOI: 10.1016/J.ESWA.2023.120489.

[20] Ullah, Fida, Gelbukh, Alexander, Zamir, Muhammad Tayyab, Riverón, Edgardo Manuel Felipe, and Sidorov, Grigori. "Enhancement of Named Entity Recognition in Low-Resource Languages with Data Augmentation and BERT Models: A Case Study on Urdu". In: *Computers 13*, no. 10 (2024), p. 258. DOI: 10.3390/COMPUTERS13100258.

[21] Ullah, Fida, Ullah, Ihsan, and Kolesnikova, Olga. "Urdu named entity recognition with attention bi-lstm-crf model". In: *Mexican International Conference on Artificial Intelligence*. Springer. 2022, pp. 3–17. DOI: 10.1007/978-3-031-19496-2_1.

[22] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*, ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[23] Weijers, Ruben and Bloem, Jelke. "An evaluation of Named Entity Recognition tools for detecting person names in philosophical text". In: *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*. 2025, pp. 418–425. DOI: 10.18653/v1/2025.nlp4dh-1.36.

[24] Wilkens, Matthew, Evans, Elizabeth F, Soni, Sandeep, Bamman, David, and Piper, Andrew. "Small Worlds: Measuring the Mobility of Characters in English-Language Fiction". In: *Journal of computational literary studies* (2024).