Automating Semantic Annotation in Low-Resource Languages: Evaluating GPT-4 for Urdu NLP

Gohar Rahman (Masaryk University)

gohartmg(at)gmail.com

Abstract

Semantic annotation is a fundamental yet labor-intensive process essential for building effective Natural Language Processing systems, particularly for low-resource languages such as Urdu. The limited availability of large, manually annotated datasets has constrained advancements in Urdu NLP. This study explores the potential of automating semantic annotation using GPT-4, a state-of-the-art large language model, through structured prompt engineering without task-specific fine-tuning. A corpus of 50,000 Urdu sentences spanning news articles, social media posts, and literary texts was used to evaluate three core tasks: Named Entity Recognition, semantic similarity, and sentiment analysis. GPT-4 demonstrated strong performance, achieving an F1-score of 92% for NER, a Pearson correlation of 0.87 for semantic similarity, and an accuracy of 88% with a macro-F1 of 87% for sentiment classification. These results indicate that LLMs guided by instruction-based prompts can reliably perform complex NLP tasks in low-resource contexts. Nonetheless, challenges with idiomatic expressions, sarcasm, and rare entities highlight the need for carefully designed prompts and potential human-AI collaboration.

Keywords

GPT-4, semantic annotation, Urdu NLP, low-resource languages, prompt engineering, Named Entity Recognition, sentiment analysis, semantic similarity

1 Introduction

Semantic annotation, the process of labeling text with meaningful metadata, is a foundational step in Natural Language Processing (NLP) pipelines. It enables machines to understand and process human language effectively, supporting tasks such as information extraction, question answering, and sentiment analysis (Navigli 2009). However, many languages remain underrepresented in NLP research, often referred to as low-resource languages. Urdu is an Indo-Aryan language spoken primarily in Pakistan and India, with large diasporic communities in the United Kingdom, the Middle East, and North America. It is written in the Perso-Arabic Nastaliq script and exhibits rich morphology, relatively flexible word order, and an elaborate system of honorifics and politeness markers, all of which present challenges for automatic language processing (Masica 1991; Schmidt 2005). NLP research in Urdu is limited due to scarce annotated corpora, inconsistent orthographic standards, and complex linguistic features (Masica 1991; Schmidt 2005).

Traditional machine learning and rule-based approaches require large amounts of labeled data, making them difficult to apply in low-resource settings.

AILing

Rahman, Gohar. 2026.
Automating semantic annotation in low-resource languages
Special Issue: *Natural Language and AI*. Vol.3 No.1
DOI: 10.62408/ai-ling.v3i1.40

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses ISSN: 2943-0070

Recent advancements in large language models (LLMs), particularly OpenAI's GPT-4 (https://openai.com/research/gpt-4), offer new possibilities for automating semantic annotation. In this study, we specifically used the GPT-4-Turbo model (gpt-4-1106-preview) accessed via the OpenAI API, which demonstrates strong performance in understanding, generating, and reasoning over natural language across multiple languages (OpenAI 2023).

LLMs can be adapted to downstream tasks through instruction-based and few-shot prompting without task-specific fine-tuning (Brown et al. 2020). Studies have shown that such models can effectively perform tasks including Named Entity Recognition (NER), sentiment analysis, and semantic similarity assessment in various languages (Shafi et al. 2022; Arif et al. 2024; Nguyen et al. 2024).

This study investigates the ability of GPT-4 to automate semantic annotation for Urdu. Specifically, it explores the application of structured prompt engineering to generate annotations for NER, semantic similarity, and sentiment classification tasks. The research aims to demonstrate that LLMs like GPT-4 can provide scalable and cost-effective solutions for enhancing NLP resources in underrepresented languages, paving the way for more comprehensive computational linguistics research in Urdu.

2 Previous research

2.1 Semantic annotation and its role in NLP

Semantic annotation, the process of enriching text with meaning-bearing labels such as entities, semantic roles, sentiment polarity, or ontology classes is foundational to many NLP applications, including question answering, information retrieval, machine translation, and sentiment analysis (Navigli 2009). Navigli's comprehensive survey underscores how semantic annotation bridges the gap between raw text and structured knowledge, enabling more precise and contextually aware NLP systems (Navigli 2009). Earlier work in the field mainly relied on rule-based approaches built upon lexicons and handcrafted grammars. While these systems offered interpretability and control, they proved fragile and difficult to generalize across diverse linguistic inputs. More recent work on semantic annotation and evaluation includes Palmer et al. (2005), Pustejovsky et al. (2019), and Devlin et al. (2019), which demonstrate the role of large-scale annotation and contextualized language models in semantic role labeling and representation learning. The rise of statistical and supervised machine learning models transformed the domain, offering greater adaptability through data-driven learning but at the cost of requiring large, annotated training datasets (Devlin et al. 2019).

Recently, the advent of transformer-based deep learning models has pushed performance further across tasks; however, even these state-of-the-art architectures continue to depend heavily on the availability of high-quality labeled data. In sum, the literature consistently positions semantic annotation as essential for effective

NLP pipelines, while also highlighting the challenges posed by resource demands and model reliance on annotated corpora.

2.2 Challenges in semantic annotation for low-resource languages

Pashto and Swahili are cited as representative low-resource languages because they have been examined in recent studies on resource scarcity and annotation challenges (e.g., Joshi et al. 2020; Nekoto et al. 2020), which discuss infrastructural and corpus limitations similar to those faced by Urdu. This scarcity is exacerbated for Urdu, which uses both a complex Perso-Arabic script (Nastaliq) and a Romanized form, each presenting distinct preprocessing and tokenization challenges (Anam et al., 2024). To date, Urdu NLP has often relied on hand-annotated datasets that are small in scale (e.g., Mukund et al. 2010; Jawaid et al. 2014; Rehman et al. 2019). However, innovative methods are emerging (Shafi et al. 2022): for instance, Anam et al. (2024) present a deep learning approach utilizing FastText and Floret embeddings, coupled with BiLSTM+GRU architectures, to significantly improve Urdu NER achieving an F-score of up to 0.98 and demonstrating the potential of neural annotation under resource constraints.

Yet such approaches still require annotated datasets for training. In sentiment analysis, the SentiUrdu-1M dataset consisting of over 1.14 million Urdu tweets annotated using weak supervision based on emoticons and sentiment lexicons represents a pivotal resource enabling large-scale experimentation even without manual labeling (Muhammad & Burney 2023). Such innovations indicate that while data scarcity remains a key challenge, creative approaches like weak supervision and embedding-based models offer viable paths forward in low-resource contexts.

2.3 Emergence of LLMs and instruction-based learning

The advent of LLMs such as GPT-3 and GPT-4 has significantly transformed NLP by enabling models to perform a wide range of tasks through natural language instructions rather than requiring task-specific training or explicit retraining (Brown et al. 2020).

This few-shot or zero-shot learning paradigm is particularly promising for low-resource languages, where annotated datasets are scarce. LLMs, pre-trained on extensive multilingual corpora, have demonstrated surprising cross-lingual capabilities, sometimes approximating human-level judgments in tasks like semantic similarity and sentiment classification even in underrepresented languages (Khurana et al. 2024).

However, performance tends to vary; biases embedded in training corpora and limited token coverage for less common languages often hinder accuracy. To improve results, researchers have developed advanced prompting techniques including cross-lingual scaffolding, translation chains, and retrieval-augmented

examples which provide contextual reinforcement from high-resource languages and thus help LLMs better generalize in low-resource settings (Khurana et al. 2024). While GPT-4's documentation notes its structured prompting ability, particularly in domains lacking extensive fine-tuning options, these techniques have yet to be thoroughly evaluated in languages like Urdu (Arif et al. 2024). Nonetheless, they offer a compelling foundation for building annotation systems that leverage LLM adaptability without requiring expensive custom models.

2.4 Prior work in Urdu NLP: NER, sentiment, and beyond

Research on Urdu NLP has made substantive strides across tasks such as NER, sentiment analysis, and corpus development. The work by Anam et al. (2024) is particularly noteworthy they achieved near state-of-the-art NER performance using deep neural architectures and embedding-based feature extraction. In sentiment analysis, Muhammad and Burney (2023) introduced the SentiUrdu-1M dataset, enabling training on massive tweet collections using automatic labeling techniques. Additionally, charged or nuanced linguistic forms such as sarcasm have received attention; datasets like the "Urdu Sarcastic Tweets Dataset" offer nearly 20,000 manually annotated utterances to support research in sarcasm detection (Khan et al. 2024). Similarly, studies addressing Roman Urdu a non-standard but widely used form include sentiment datasets and hybrid embedding approaches; for example, Memood et al. (2020) developed and evaluated word embeddings trained specifically on Roman Urdu, using traditional ML and deep learning models to achieve strong baseline performance (Ahmad & Jundran, 2025; Shafi et al. 2022). Recent work shows that large multilingual language models can perform semantic annotation tasks in low-resource settings using zero-shot and few-shot prompting (Lauscher et al. 2020; Lin et al. 2022). While these contributions mark significant progress, gaps persist particularly in building unified annotation systems across tasks using minimal supervision or flexible architectures suited for low-resource deployment.

2.5 Synthesis and research gap

The literature collectively underscores two truths: first, semantic annotation is indispensable for enabling robust NLP pipelines; second, the primary obstacle for low-resource languages like Urdu is the scarcity of labeled data and language-specific resources. Existing work demonstrates the effectiveness of deep learning models (Anam et al. 2024), weak supervision (Muhammad & Burney 2023)), and dedicated datasets for sarcasm or Romanized forms (Khan et al.2022; Memood et al. 2020). Yet these methods generally focus on isolated tasks. Meanwhile, the rise of instruction-based learning using LLMs creates opportunities to unify annotation tasks such as NER, semantic similarity, and sentiment classification under a single framework that requires only prompt construction rather than retraining. To date,

such an instruction-based, multi-task annotation approach for Urdu remains unexplored. This gap forms the motivation for the present study: evaluating whether GPT-4, guided by structured prompt engineering, can reliably automate semantic annotation across multiple tasks in Urdu, thereby offering a scalable and cost-effective alternative to traditional pipelines.

3 Data and procedures

This study investigates the application of GPT-4 for automating semantic annotation in Urdu. The methodology involves corpus collection, task definition, and the use of prompt engineering and structured instruction-based workflows to generate annotations. To ensure transparency and replicability, detailed corpus examples, prompt templates, annotation guidelines, and evaluation metrics are documented in Appendix A.

3.1 Data collection and evaluation design

A diverse Urdu corpus was compiled to represent multiple domains, including news articles, social media posts, and literary texts. The corpus consisted of 50,000 sentences, ensuring coverage of formal, informal, and colloquial language. Given that Urdu is a low-resource language, special attention was paid to collecting data that captured linguistic variability, including variations in spelling, Roman Urdu, and script differences (Soomro et al. 2024).

As detailed in Appendix A (Writing System and Transliteration), Nastaliq script was retained for formal sources such as news and literary texts, while Roman Urdu was preserved for social media and informal communication. Both writing systems were included in the evaluation subsets to enable comparison of model performance across scripts. No automatic transliteration was applied unless explicitly stated, and the original scripts were maintained in GPT-4 prompts to reflect authentic language use (Appendix A.1).

The collected data were preprocessed to remove noise such as duplicate entries, incomplete sentences, and encoding errors. Preprocessing steps included Unicode normalization, removal of optional diacritics, standardization of character variants (e.g., different forms of Yeh and Heh), token normalization, and rule-based harmonization of common Roman Urdu spelling variants (Appendix A.1).

The complete 50,000-sentence corpus was used as a linguistic pool for prompt construction, contextual diversity, and large-scale automatic annotation. For quantitative evaluation, smaller gold-standard subsets were created through manual annotation: 500 sentences for NER, 500 sentence pairs for semantic similarity, and 1,000 sentences for sentiment analysis. Sentence extraction followed a semi-automatic procedure. Source texts were first collected in full, after which sentence boundaries were identified using punctuation-based rules. The resulting sentences were then manually reviewed to ensure correctness and completeness.

GPT-4 outputs on these subsets were compared against human annotations to compute performance metrics. It should be noted that short evaluative sentences such as *Yeh restaurant bohat acha hai* reflect naturally occurring language commonly found in social media and conversational contexts. Such examples were intentionally included to capture informal and user-generated language relevant for sentiment analysis. Thus, the large corpus ensured representativeness, while the smaller subsets enabled controlled and reliable evaluation (Appendix A.6).

The 50,000-sentence Urdu corpus was compiled from publicly accessible Urdu-language sources across three domains: (i) news articles, (ii) social media posts, and (iii) literary texts. News data were collected from widely used Urdu news platforms such as Dawn Urdu, BBC Urdu, Express News, and Jang. Social media data were sampled from publicly available posts and comments on platforms such as X (formerly Twitter) and Facebook. Literary texts were drawn from open-access digital Urdu repositories and online prose collections. Due to copyright and platform restrictions, the full corpus cannot be publicly redistributed. However, the data were collected from publicly accessible sources, and representative samples along with annotation procedures are provided to ensure transparency and reproducibility.

Sentence extraction was conducted using a semi-automatic approach. Texts were first collected at the document level, and sentence segmentation was performed using rule-based and punctuation-based splitting. This was followed by manual verification to ensure accuracy and remove noise.

As the dataset was constructed from publicly available materials for research purposes, individual sentence-level URLs were not systematically retained. The full corpus is therefore not publicly redistributed; however, representative examples and detailed annotation procedures are provided in Appendix A to ensure transparency and reproducibility.

Sentences were selected based on the following criteria: (a) grammatical completeness, defined in this study as the presence of a well-formed syntactic structure containing at least a subject–predicate relationship or a complete clause expressing a meaningful proposition, (b) relevance to everyday or public discourse, and (c) representation of both Nastaliq and Roman Urdu writing systems. Duplicate content, incomplete fragments, headlines without predicates, advertisements, and non-Urdu material were excluded during preprocessing.

Domain categorization (news, social media, literary) was carried out manually at the source level, based on the origin and communicative function of the text, prior to sentence-level extraction. This ensured consistent categorization while maintaining linguistic diversity across domains. A summary of corpus composition is provided in Appendix A.6.

3.2 Semantic annotation tasks

The study focused on three semantic annotation tasks to evaluate GPT-4's performance in Urdu. The first task was NER, which involves identifying and classifying named entities into predefined categories such as PERSON, ORGANIZATION, LOCATION, and DATE. In the present annotation scheme, personal titles and institutional roles (e.g., وزیر اعظم 'prime minister', صدر 'president', چیف جسٹس 'chief justice') were consistently annotated as PERSON entities, following common practice in NER guidelines (Nadeau & Sekine 2007). NER is a foundational semantic annotation task supporting information extraction and knowledge base construction (Nadeau & Sekine 2007). Sample NER prompts and annotated examples are provided in Appendix A.2.1 and Appendix A.4.In the NER annotation scheme, each sentence is processed by identifying spans of text corresponding to named entities. Each identified entity is then assigned one of four predefined labels: PERSON, ORGANIZATION, LOCATION, or DATE. Annotation is span-based, meaning that contiguous tokens forming a single entity (e.g., وزیر اعظم) are labeled as one unit. The output is structured in JSON format, where each entity is represented by its textual span and assigned category label. For example, in the sentence *وزیر اعظم نے اسلام آباد میں نئی پالیسی کا اعلان کیا*, the phrase *وزیر اعظم* is annotated as PERSON and *اسلام آباد* as LOCATION. The examples illustrate how annotation is applied step-by-step, including identification of entity spans, assignment of labels, and structured output formatting.

The second task was semantic similarity, which measures the degree of semantic equivalence between sentence pairs. This task is critical for applications such as paraphrase detection, question answering, and information retrieval (Arif et al. 2024). Prompt design and example sentence pairs are illustrated in Appendix A.2.2, with annotated samples in Appendix A.4.The third task was sentiment analysis, which classifies text as positive, negative, or neutral and is widely used to assess public opinion and user feedback. Example prompts and labeled outputs are documented in Appendix A.2.3 and Appendix A.4.

3.3 GPT-4 integration using prompt engineering

As GPT-4 cannot be fine-tuned in a traditional supervised manner, prompt engineering was employed to guide the model's behavior. Prompt construction followed a consistent structure with three components, as outlined in Appendix A.5.

First, each prompt included explicit task instructions in Urdu or in a bilingual Urdu-English format (e.g., "Identify all named entities in the following Urdu sentence"). The bilingual formulation was adopted to leverage GPT-4's stronger instruction-following capabilities in English while simultaneously anchoring the task in the target language, Urdu. This design was intended to improve output consistency and label accuracy by reducing potential ambiguity in

purely Urdu-only instructions and by aligning the model's internal task representations with well-established English NLP terminology. Second, few-shot examples were provided to demonstrate the expected output format, leveraging instruction-based learning (Brown et al. 2020). Third, structured output constraints (e.g., JSON or tables) were specified to ensure machine-readable results suitable for downstream processing. Full prompt templates for each task are presented in Appendix A.2.

For large-scale annotation, the corpus was processed in batches of 100 sentences per API call, as detailed in Appendix A.7. Each sentence was treated as an independent input item within the batch, and no conversational or cross-sentence context was provided to the model; that is, the prompts were structured so that GPT-4 generated annotations for each sentence separately, without access to preceding or following sentences, in order to avoid topic or entity carryover effects.

The resulting GPT-4 outputs were automatically post-processed using Python scripts to validate structural well-formedness (e.g., correct JSON syntax, presence of required fields, and label conformity to the predefined tag set) and to normalize formatting. This validation concerned output format consistency rather than semantic correctness. A subset of the automatically generated annotations was then compared against manually annotated gold-standard data to assess accuracy using the evaluation metrics reported in Section 3.4 (OpenAI 2023).

For NER and sentiment analysis, single-sentence inputs were used. For semantic similarity, both sentences in each pair were provided together without additional discourse context to avoid topical bias. Few-shot examples were limited to two to three instances per prompt to balance performance and token efficiency, following the guidelines summarized in Appendix A.5.

Sentence pairs were constructed using a mixed strategy. One sentence in each pair was sampled from the corpus, while the second sentence was either (a) a manually created paraphrase, (b) a semantically related sentence from the corpus, or (c) an unrelated sentence. This approach yielded high-, medium-, and low-similarity pairs, enabling graded evaluation of GPT-4's similarity judgments. Example sentence pairs and corresponding similarity scores are shown in Appendix A.4.

3.4 Evaluation

The automated annotations were evaluated using standard NLP metrics. For NER, Automated annotations were evaluated using standard NLP metrics, with detailed definitions provided in Appendix A.3. For NER, precision, recall, and F1-score were computed. Semantic similarity was evaluated by comparing GPT-4–generated similarity scores (ranging from 0 to 1) with human-annotated similarity ratings using Pearson correlation.

Sentence embeddings were computed using the OpenAI *text-embedding-3-large* model. These embeddings served as an external semantic baseline, and cosine

similarity was used to quantify vector-space proximity between sentence pairs for correlation analysis with GPT-4 similarity scores and human judgments.Cosine similarity is meaningful in this context because it provides a continuous measure of semantic closeness in the same [0,1] range, allowing direct comparison with human and GPT-4 similarity scores.

To establish a reliable human baseline, three trained Urdu linguistics graduates independently annotated the evaluation subsets using the same task instructions supplied to GPT-4 (see Appendix A.2). Inter-annotator agreement was calculated using Cohen's Kappa (McHugh 2012) for the categorical tasks, yielding $\kappa = 0.87$ for NER and $\kappa = 0.82$ for sentiment analysis. Annotators were provided with written task-specific guidelines for each annotation task (NER, sentiment analysis, and semantic similarity), including label definitions, decision rules, and illustrative examples (see Appendix A.2). Guidelines were approximately 3–4 pages in length and were discussed in a short training session prior to annotation.

Annotation was conducted independently. Annotators did not have access to each other's annotations and did not see GPT-4 outputs at any stage of the annotation process. In cases of disagreement, no automatic adjudication was performed for agreement calculation; instead, disagreements were resolved through majority voting to produce the final gold-standard labels used for evaluation. For the semantic similarity task, which involved continuous ratings on a 0–1 scale, agreement was assessed using the Intraclass Correlation Coefficient (ICC), resulting in an ICC of 0.79, indicating substantial consistency among annotators. To establish a reliable human baseline, three trained Urdu linguistics graduates independently annotated the evaluation subsets using the same task instructions supplied to GPT-4 (see Appendix A.2). For the categorical tasks NER and sentiment analysis inter-annotator agreement was computed using Cohen's Kappa (McHugh 2012), yielding $\kappa = 0.87$ for NER and $\kappa = 0.82$ for sentiment analysis.

For the semantic similarity task, annotators provided continuous similarity ratings on a 0–1 scale. As Cohen's Kappa is designed for categorical data, it was not used for this task. Instead, inter-annotator agreement was assessed using the Intraclass Correlation Coefficient (ICC), which is appropriate for continuous measurements. The resulting ICC value of 0.79 indicates substantial agreement among annotators. No discretization or binning of similarity scores was applied, as the study aimed to preserve the continuous nature of semantic similarity judgments.

3.5 Challenges and mitigation strategies

Several challenges were encountered during the annotation process. One challenge was orthographic ambiguity in Urdu, including homographs and spelling variation, which was mitigated through preprocessing and normalization (Appendix A.1). Another challenge was context sensitivity, as GPT-4's outputs can vary depending on prompt framing; this was addressed through carefully designed task instructions and few-shot examples (Appendix A.5). Finally, domain variability posed

challenges due to differences in vocabulary and style across news, social media, and literary texts. To enhance robustness, representative examples from each domain were included in the prompts, as illustrated in Appendix A.1 and A.2.

4 Results

The performance of GPT-4 in automating semantic annotation for Urdu was evaluated across three primary tasks: NER, semantic similarity, and sentiment analysis. The evaluation metrics demonstrate GPT-4's ability to generate high-quality annotations using prompt engineering and structured instruction-based workflows. The results for each task are presented below.

4.1 NER

The model's performance in NER was evaluated using precision, recall, and F1-score on a manually annotated test set. GPT-4 achieved a precision of 91%, indicating that most of the entities identified by the model were correct. The recall score of 93% shows that the model successfully captured the majority of true entities present in the corpus. The resulting F1-score of 92% reflects a strong balance between precision and recall, demonstrating consistent and reliable performance.

These results indicate that GPT-4 is effective at recognizing named entities across diverse Urdu text domains, including news articles, social media posts, and literary texts. The model correctly identified common entity types such as locations (e.g., *Islamabad*), persons (e.g., *Wazir-e-Azam* [Prime Minister]), and organizations (e.g., *Pakistan Cricket Board*).

Nevertheless, some limitations were observed. GPT-4 occasionally misclassified domain-specific or less frequent entities, particularly titles of literary works and culturally specific product names. For instance, *Shahab Nama* the title of a well-known Urdu autobiographical book by Qudratullah Shahab was sometimes labeled as a PERSON rather than as a creative work, reflecting the model's tendency to associate capitalized or name-like expressions with human entities. Similarly, *Rooh Afza*, a popular South Asian herbal beverage brand, was occasionally misclassified as a LOCATION or ORGANIZATION instead of a product entity. These errors indicate that, in the absence of fine-grained domain adaptation, GPT-4 relies primarily on surface lexical cues and distributional prominence, which can lead to ambiguity when proper nouns denote cultural artifacts rather than persons or institutions. These errors likely stem from the absence of task-specific fine-tuning and the reliance on prompt-based instruction alone. Overall, the high precision, recall, and F1-score values confirm GPT-4's robustness for Urdu NER, while also highlighting areas where domain adaptation or supplementary annotation guidelines could further improve performance.

4.2 Semantic similarity

Semantic similarity was evaluated using 500 sentence pairs, with GPT-4's similarity judgments compared against human-annotated ratings. The model's scalar similarity scores (0–1) were compared with vector-based semantic similarity obtained from sentence embeddings generated using the OpenAI *text-embedding-3-large* model. Cosine similarity between these reference embeddings yielded an average score of 0.89, and the correlation between GPT-4 similarity scores and human judgments reached 0.87 (Pearson's r), indicating strong alignment between distributional semantic proximity, model predictions, and human perception. These results indicate a strong alignment between GPT-4 predictions and human semantic judgments.

Overall, GPT-4 demonstrated a robust ability to capture semantic equivalence in Roman Urdu sentences, even in cases involving lexical variation or paraphrasing. For example, the sentence pair *Woh school ja rahahai* ('He is going to school') and *Woh taleem keliye ja rahahai* ('He is going to attend classes / to study') received high similarity scores, as both express the same underlying event of going somewhere for the purpose of education, despite lexical variation in the surface forms. In contrast, pairs with reduced informational overlap such as *Woh cricket khel raha hai* ('He is playing cricket') versus *Woh khel raha hai* ('He is playing') were assigned moderate similarity scores due to loss of specificity. Sentences with divergent meanings, such as *Aaj Mausam khushgawar hai* ('The weather is pleasant today') and *Aaj barish hone walihai* ('It is going to rain today'), were correctly assigned low similarity scores despite topical overlap.

While GPT-4 generally aligned closely with human judgments, limitations were observed for culturally specific idiomatic expressions. For instance, the idiom اونٹ کے منہ میں زیرہ (*Oont ke munh mein zeera*) 'A drop in the ocean' was paired with the sentence یہ فائدہ بہت کم ہے (*Ye faida bohot kam hai*) 'This benefit is very small'. Although human annotators assigned a high similarity score (0.90), GPT-4 substantially underestimated the semantic equivalence (0.41). This discrepancy suggests that figurative and culturally grounded meanings are not always fully captured by instruction-based prompting alone.

Taken together, these findings confirm GPT-4's effectiveness in modeling semantic similarity by assigning continuous similarity scores (ranging from 0 to 1) to sentence pairs based on their semantic equivalence. These scores are generated through instruction-based prompting, where the model evaluates the degree of meaning overlap between two sentences using its contextual and distributional understanding of language.

4.3 Sentiment analysis

Sentiment classification was evaluated on a test set of 1,000 sentences. GPT-4 achieved an accuracy of 88% and a macro-F1 score of 87%, reflecting balanced

performance across positive, negative, and neutral categories, even with uneven class distribution.

These results indicate GPT-4's strong ability to detect sentiment in diverse Roman Urdu texts, including literary writing, formal news, and informal social media posts. For example:

- *Yeh restaurant bohat acha hai* ('This restaurant is very good') → correctly classified as POSITIVE.
- *Aaj match bohat bura tha* ('The match was very bad today') → correctly classified as NEGATIVE.
- *Woh school gaya* ('He went to school') → correctly classified as NEUTRAL.

To further illustrate challenging cases, consider the following examples from the corpus:

- *Film achi thi lekin bohat lambi thi* ('The movie was good but very long') → mixed sentiment (POSITIVE + NEGATIVE)
- *Wah kya shandar service hai!* ('Wow, what excellent service!') → Sarcastic (NEGATIVE intended)
- *Theek tha, lekin zyada khaas nahi* ('It was okay, but nothing special') → SUBTLE/NEUTRAL-NEGATIVE

The following examples are representative sentences drawn from the compiled corpus and are presented in anonymized form. As the dataset was constructed from publicly available but non-redistributable sources, individual sentence-level URLs are not provided.

Misclassifications in sentiment analysis arise primarily from linguistic and contextual complexity rather than simple polarity detection. In particular, sentences containing mixed sentiment (e.g., positive and negative elements within the same utterance), sarcasm, or implicit evaluation pose challenges for automated systems. For example, a sentence such as *Film achi thi lekin bohat lambi thi* ('The movie was good but very long') contains both positive and negative cues, requiring the model to determine overall sentiment orientation. Similarly, sarcastic expressions such as *Wah kya shandar khidmat hai!* ('Wow, what excellent service!' used ironically) rely on contextual and pragmatic interpretation that may not be fully captured by instruction-based prompting. These factors contribute to occasional misclassification despite otherwise strong overall performance.

A comparison between Nastaliq and Roman Urdu subsets revealed minor variation in performance. GPT-4 showed slightly higher consistency on Roman Urdu data, likely due to greater representation of Latin-script text in its training data. However, overall performance remained robust across both writing systems. While examples in Section 4.3 are presented in Roman Urdu for readability, the evaluation dataset included both scripts.

Table 1: Category-wise precision, recall and F1-scores for GPT-4 sentiment classification.

Sentiment	Precision	Recall	F1-score
Positive	0.89	0.91	0.90
Negative	0.86	0.85	0.85
Neutral	0.88	0.87	0.87

Table 1 presents the category-wise performance of GPT-4 on the Urdu sentiment classification task. The model achieved the highest F1-score for the positive class (0.90), followed closely by neutral (0.87) and negative (0.85) categories. The relatively balanced precision and recall values across all three classes indicate that GPT-4 does not exhibit strong bias toward any single sentiment category and is able to reliably distinguish between positive, negative, and neutral expressions in Urdu text.

These results provide a more fine-grained view of model performance than aggregate accuracy and support the claim that GPT-4 performs consistently across different affective polarities.

4.4 Summary of results

Overall, GPT-4 demonstrated robust and consistent performance across all evaluated semantic annotation tasks, underscoring its potential for large-scale annotation in Urdu and Roman Urdu. In the NER task, the model achieved the strongest results, with an F1-score of 92%, reflecting its ability to accurately and consistently identify entities such as persons, locations, and organizations across multiple domains.

For semantic similarity, GPT-4's predictions showed a high degree of alignment with human judgments, achieving a Pearson correlation of 0.87 and a cosine similarity score of 0.89. These results indicate that the model effectively captures semantic equivalence even in cases involving paraphrasing or lexical variation. In sentiment analysis, GPT-4 attained an accuracy of 88% and a macro-F1 score of 87%, demonstrating reliable and balanced performance across positive, negative, and neutral sentiment classes, including informal and social media-style text.

These findings confirm the effectiveness of instruction-based semantic annotation using GPT-4 for Urdu, a traditionally low-resource language. By reducing reliance on extensive manual annotation, this approach offers a scalable and cost-efficient alternative for building annotated datasets for downstream NLP tasks. At the same time, certain limitations were observed, particularly in handling culturally specific idioms, sarcasm, and rare or domain-specific entities, suggesting avenues for future work in domain adaptation and hybrid annotation strategies.

Regarding script variation, approximately 82% of the corpus was in Nastaliq script and 18% in Roman Urdu. Performance was marginally higher on Roman Urdu texts, likely reflecting the model's greater exposure to Latin-script data during pretraining. This observation highlights the importance of script-aware prompt design and balanced corpus composition when applying LLMs to Urdu.

5 Discussion

The findings of this study provide compelling evidence that GPT-4, when carefully guided through structured prompts and instruction-based workflows, is capable of performing semantic annotation tasks in Roman Urdu with a high degree of reliability. This is particularly significant given that Urdu, despite being the national language of Pakistan and spoken by millions worldwide, remains underrepresented in computational linguistics research and under-resourced in terms of annotated corpora. The results suggest that LLMs, even without task-specific fine-tuning, can extend their general linguistic competencies into low-resource languages, thereby challenging the longstanding assumption that progress in NLP requires labor-intensive annotation of large datasets. In what follows, the discussion interprets the model's performance across three major tasks (NER, semantic similarity, and sentiment analysis) while also situating these findings in relation to prior work, and drawing out their wider implications for NLP in low-resource contexts.

One of the most striking outcomes of this study was the strong performance of GPT-4 in NER. Achieving an F1-score of 92%, the model showed remarkable proficiency in identifying and classifying named entities such as persons, organizations, and locations. This finding aligns closely with Brown et al. (2020), who demonstrated that large pretrained models such as GPT-3 could, without explicit fine-tuning, perform well on entity-level tasks when supplied with structured prompts and relevant examples. Similarly, Arif et al. (2024) have argued that instruction-based prompting frameworks allow even general-purpose LLMs to adapt effectively to low-resource annotation challenges, a claim that is substantiated by the present results in the Urdu context. This underlines the importance of prompt design and corpus diversity: prompts must include varied examples in order to encourage the model to generalize beyond frequent categories. The overall success of GPT-4 in NER therefore demonstrates its potential as a cost-efficient alternative to traditional rule-based and supervised approaches, which often require extensive domain expertise and labeled data that are scarce in Urdu and other underrepresented languages.

The performance of GPT-4 in semantic similarity tasks further underscores the model's capability to generalize semantic understanding in a low-resource setting. With a cosine similarity of 0.89 computed between sentence embeddings generated using the OpenAI *text-embedding-3-large* model and a Pearson correlation of 0.87 against human judgments, GPT-4 demonstrated its ability to approximate human-like evaluations of semantic equivalence even in the presence

of lexical variation, paraphrasing, or synonymy. These results reinforce earlier findings by Nguyen et al. (2024), who showed that LLMs could capture semantic equivalence across linguistically diverse sentence structures, and by Arif et al. (2024), who noted the strength of instruction-based workflows in supporting semantic alignment. Importantly, the occasional challenges with idiomatic or culturally specific expressions echo the limitations that cultural grounding remains a major barrier for NLP models in underrepresented languages. This suggests that while GPT-4 performs strongly overall, its semantic similarity judgments may benefit from the inclusion of culturally contextualized prompts, or even hybrid approaches that combine statistical similarity measures with culturally informed rules. In terms of practical applications, the ability to capture semantic similarity has wide-ranging implications, from plagiarism detection and information retrieval to question answering and automatic evaluation of translations. The results here thus extend the insights of Nguyen et al. (2024) by showing that these capacities are not limited to high-resource languages but can also be harnessed for Roman Urdu, thereby broadening the accessibility of advanced NLP tools.

Sentiment analysis presented another area where GPT-4 displayed consistent and competitive performance. Achieving an accuracy of 88% and a macro-F1 of 87%, the model demonstrated robustness in classifying text into positive, negative, and neutral categories. This is particularly noteworthy given the complexities of sentiment detection in Urdu, where expressions often include mixed emotions, sarcasm, or irony. The results resonate with Soomro (2024), who emphasized the inherent difficulty of sentiment classification in Urdu due to informal registers, colloquial variation, and the use of sarcasm in online communication. However, the present findings go further by showing that prompt engineering particularly through the use of explicit task instructions and few-shot examples is sufficient to guide GPT-4 toward near-human performance. This supports the broader conclusions of Brown et al. (2020) and OpenAI (2023), who observed that LLMs can achieve strong results in classification and annotation tasks when structured prompting is employed, even in the absence of fine-tuning. Addressing this may require hybrid workflows in which GPT-4 serves as the first-pass annotator, while human validators review ambiguous or nuanced cases. Such a workflow would not only improve reliability but also maintain scalability, which is one of the key strengths of instruction-based annotation.

Beyond task-specific findings, the broader implications of this study are significant for the future of NLP in low-resource languages. First, the scalability of GPT-4 provides a practical solution to the longstanding bottleneck of data scarcity. As Navigli (2009) argued more than a decade ago, the lack of annotated corpora has been a critical obstacle to the development of NLP resources for many languages. The present results demonstrate that GPT-4 can generate high-quality annotations at scale, thereby reducing reliance on costly manual labeling and accelerating resource creation for Urdu. Second, the adaptability of GPT-4 through prompt engineering is consistent with the conclusions of Arif et al. (2024), who

highlighted the versatility of instruction-based workflows across multiple annotation tasks. Rather than retraining models for each individual task, practitioners can now design task-specific prompts that unlock the latent knowledge of the model, making LLMs highly flexible for underrepresented contexts. Third, the quality of the outputs observed here competitive with traditional supervised models despite the lack of fine-tuning provides empirical support for the claims of Brown et al. (2020) and OpenAI (2023), who noted the surprising generalization capabilities of LLMs across domains. Finally, however, the limitations of GPT-4 observed in this study reinforce the concerns raised by (Nadeau & Sekine 2007), who cautioned against assuming that LLMs are universally reliable across cultural and linguistic domains. Challenges with idiomatic interpretation, rare entities, and implicit sentiment detection underscore the importance of combining automated annotation with human oversight and of designing culturally grounded prompts.

Taken together, these findings indicate that GPT-4 and similar LLMs can play a transformative role in NLP for low-resource languages like Urdu. They confirm the scalability, adaptability, and quality of instruction-based workflows while also highlighting areas for refinement and further research. Previous Urdu NER and sentiment studies based on CRF and BiLSTM models report F1-scores in the range of 0.70–0.85 (Mukund et al., 2010; Riaz 2010). The higher performance observed in the present study suggests that LLMs offer substantial gains for semantic annotation in low-resource settings. Moreover, the study contributes to broader debates about linguistic equity in NLP, as outlined by Navigli (2009). By showing that a state-of-the-art LLM can achieve near-human annotation performance in Urdu, the results provide a counterpoint to the dominance of English and other high-resource languages in NLP research. At the same time, they raise critical questions about sustainability and accessibility: while proprietary models such as GPT-4 offer immediate benefits, the long-term development of NLP for underrepresented languages will also require open-source initiatives, culturally aware evaluation frameworks, and domain-specific fine-tuning strategies. In this sense, GPT-4 can be understood not as a replacement for traditional corpus-building but as a catalyst that accelerates the creation of resources and enables researchers to focus human expertise where it is most neededon the culturally specific and nuanced aspects of language that remain challenging for LLMs.

6 Conclusion

This study highlights the feasibility and practical effectiveness of employing GPT-4 for the automation of semantic annotation in Urdu, a language that has traditionally been considered low-resource within the domain of NLP. Through the application of prompt engineering and carefully structured instruction-based workflows, GPT-4 was able to undertake complex annotation tasks such as NER, semantic similarity assessment, and sentiment analysis with commendable accuracy. These findings demonstrate that LLMs, when guided by explicit task

instructions, can offer scalable and cost-efficient alternatives to conventional annotation methods, thereby alleviating the need for extensive manual labeling, which has often posed a barrier to the creation of large and reliable datasets in low-resource contexts. The success across multiple tasks suggests that GPT-4 can not only replicate but also streamline processes that are typically labor-intensive, accelerating the pace of resource development for languages like Urdu.

However, the results also bring to light certain limitations that require consideration for future work. GPT-4 occasionally produced errors in the recognition of rare or domain-specific entities, revealed sensitivity to idiomatic expressions during semantic similarity evaluation, and demonstrated difficulty in accurately identifying nuanced or mixed sentiments. These challenges emphasize the need for careful prompt formulation and the inclusion of diverse, representative examples within task design. They also point to the potential value of incorporating human-in-the-loop validation, particularly in high-stakes applications where precision is critical. Such hybrid strategies could balance the efficiency of automated annotation with the reliability of human oversight, ensuring both scalability and contextual sensitivity.

Overall, the findings carry meaningful implications for NLP in underrepresented languages. They suggest that instruction-based adaptation of LLMs like GPT-4 could play a transformative role in advancing linguistic resource development, supporting applications such as information extraction, machine translation, and sentiment monitoring. Future research may build on this foundation by combining GPT-4 outputs with rule-based systems or fine-tuned models, as well as extending the approach to other low-resource languages to test its generalizability and adaptability across diverse linguistic landscapes

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this contribution.

Acknowledgment

The author would like to sincerely thank Izza Rashid, Hamid Ullah, and Muhammad Ibrahim for their diligent work in manually annotating the evaluation datasets. Their expertise in Urdu linguistics and careful adherence to the annotation guidelines were essential for establishing a reliable human benchmark, without which this study would not have been possible. The author also acknowledges the use of OpenAI's GPT-4-Turbo in assisting with the editing and refinement of this paper, enhancing both its clarity and presentation in English. However, the author ensured that all contributions adhered strictly to the standards and ethical guidelines of academic writing.

References

Ahmad, Muhammad & Jundran, Azim Ullah. 2025. Factors and forms of semantic change in Urdu language: An analytical study. *Negotiations* 5(1). 200–212.

Anam, Rimsha & Anwar, Muhammad Waqas & Jamal, Muhammad Hasan & Bajwa, Usama Ijaz & de la Torre Diez, Isabel & Silva Alvarado, Eduardo & Soriano Flores, Emmanuel, & Ashraf, Imran. 2024. A deep learning approach for named entity recognition in Urdu language. *PLoS ONE* 19(3), e0300725. 1–21. https://doi.org/10.1371/journal.pone.0300725

Arif, Samee & Azeemi, Abdul Hameed & Raza, Agha Ali, & Athar, Awais. 2024. Generalists vs. specialists: Evaluating large language models for Urdu. *Findings of the Association for Computational Linguistics: EMNLP 2024*. 426–435. https://doi.org/10.48550/arXiv.2407.04459

Brown, Tom B. & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind & Shyam, Pranav & Sastry, Girish & Askell, Amanda & Agarwal, Sandhini & Herbert-Voss, Ariel & Krueger, Gretchen & Henighan, Tom & Child, Rewon & Ramesh, Aditya & Ziegler, Daniel M. & Wu, Jeffrey & Winter, Clemens & Hesse, Christopher & Chen, Mark & Sigler, Eric & Litwin, Mateusz & Gray, Scott & Chess, Benjamin & Clark, Jack & Berner, Christopher & McCandlish, Sam & Radford, Alec & Sutskever, Ilya & Amodei, Dario. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* (*NeurIPS 2020*). 3–75. *arXiv.* https://arxiv.org/abs/2005.14165 (last accessed on ...).

Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT 2019*. 4171–4186. https://doi.org/10.18653/v1/n19-1423

Joshi, Pratik & Santy, Sebastian & Budhiraja, Amar & Bali, Kalika & Choudhury, Monojit. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *ACL 2020*. 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

James Pustejovsky & Ken Lai & Nianwen Xue. 2019. Modeling quantification and scope in Abstract Meaning Representations. In Nianwen Xue, William Croft, Jan Hajič, Chu-Ren Huang, Stephan Oepen, Martha Palmer, and James Pustejovsky (eds), *Proceedings of the First International Workshop on Designing Meaning Representations*, 28–33. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3303

Khan, Shumaila & Qasim, Iqbal & Khan, Wahab & Khan, Aurangzeb & Khan, Javed Ali & Qahmash, Ayman & Ghadi, Yazeed Yasin. 2024. An automated

approach to identify sarcasm in low-resource language. *PLoS ONE*, 19(12), e0307186. 1–29. https://doi.org/10.1371/journal.pone.0307186

Khurana, Sameer & Dawalatabad, Nauman & Laurent, Antoine & Vicente, Luis & Gimeno, Pablo & Mingote, Victoria. 2024. Cross-lingual transfer learning for low-resource speech translation. *IEEE ICASSP Workshops*. 670–674. https://doi.org/10.1109/icasspw62465.2024.10626683

Khan, Lal & Amjad, Ammar & Ashraf, Noman & Chang, Hsien-Tsung. 2022. Multi-class sentiment analysis of Urdu text using multilingual BERT. *Scientific Reports* 12(1). 5436. https://doi.org/10.1038/s41598-022-09381-9

Lauscher, Anne & Ravishankar, Vinit & Vulić, Ivan & Glavaš, Goran. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. *EMNLP 2020*. 4483–4499. https://doi.org/10.18653/v1/2020.emnlp-main.363

McHugh, Mary L. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3). 276–282. https://doi.org/10.11613/BM.2012.031

Memood, Faiza & Ghani, Muhammad Usman & Ibrahim, Muhammad Ali & Shehzadi, Rehab & Asim, Muhammad Nabeel. 2020. A precisely Xtreme-multi channel hybrid approach for Roman Urdu sentiment analysis. *arXiv.* https://arxiv.org/abs/2003.05443 (last accessed on...).

Mukund, Smruthi & Srihari, Rohini & Peterson, Erik. 2010. An information-extraction system for Urdu – a resource-poor language. *ACM Transactions on Asian Language Information Processing* 9(4). 1–43. https://doi.org/10.1145/1838751.1838754

Nadeau, David & Sekine, Satoshi. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes* 30(1). 3–26. https://doi.org/10.1075/li.30.1.03nad

Masica, Colin P. 1991. The Indo-Aryan languages. Cambridge University Press.

Muhammad, Khalid Bin & Burney, Syed Muhammad Aqil. 2023. Innovations in Urdu sentiment analysis using machine and deep learning techniques for two-class classification of symmetric datasets. *Symmetry* 15(5). 1027. 1–14. https://doi.org/10.3390/sym15051027

Nekoto, Wilhelmina & Marivate, Vukosi & Matsila, Tshinondiwa & Fasubaa, Timi & Fagbohungbe, Taiwo & Akinola, Solomon Oluwole & Muhammad, Shamsuddeen & Kabenamualu, Salomon Kabongo & Osei, Salomey & Sackey, Freshia & Niyongabo, Rubungo Andre & Macharm, Ricky & Ogayo, Perez & Ahia, Orevaoghene & Berhe, Musie Meressa & Adeyemi, Mofetoluwa & Mokgesi-Selinga, Masabata & Okegbemi, Lawrence & Martinus, Laura & Tajudeen, Kolawole & Degila, Kevin & Ogueji, Kelechi & Siminyu, Kathleen & Kreutzer, Julia & Webster, Jason & Toure Ali, Jamiil & Abbott, Jade & Orife, Iroro & Ezeani, Ignatius & Dangana, Idris Abdulkadir & Kamper,

Herman & Elsahar, Hady & Duru, Goodness & Kioko, Ghollah & Espoir, Murhabazi & van Biljon, Elan & Whitenack, Daniel & Onyefuluchi, Christopher & Emezue, Chris Chinenye & Dossou, Bonaventure F. P. & Sibanda, Blessing & Bassey, Blessing & Olabiyi, Ayodele & Ramkilowan, Arshath & Öktem, Alp & Akinfaderin, Adewale & Bashir, Abdallah. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In Trevor Cohn, Yulan He, and Yang Liu (eds), *Findings of the Association for Computational Linguistics: EMNLP* 2020. 2144–2160. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.195

Nguyen, Xuan-Phi & Aljunied, Sharifah Mahani & Joty, Shafiq & Bing, Lidong. 2024. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. *arXiv.* https://doi.org/10.48550/arxiv.2306.11372

Navigli, Roberto. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys* 41(2). 1–69. https://doi.org/10.1145/1459352.1459355

OpenAI. 2023. GPT-4 technical report. https://openai.com/research/gpt-4 (last accessed on…).

Palmer, Martha & Gildea, Daniel & Kingsbury, Paul. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics* 31(1). 71–106. https://doi.org/10.1162/0891201053630264

Riaz, Khuram. 2010. Rule-based named entity recognition in Urdu. *Proceedings of the 2010 Named Entities Workshop*. Uppsala: Association for Computational Linguistics. 126–135.

Shafi, Jawad & Iqbal, Hafiz Rizwan & Nawab, Rao Muhammad Adeel & Rayson, Paul. 2022. UNLT: Urdu Natural Language Toolkit. *Natural Language Engineering* 29(4) 942–977. https://doi.org/10.1017/s1351324921000425

Schmidt, Ruth Laila. 2005. *Urdu: An Essential Grammar*. London: Routledge. https://doi.org/10.4324/9780203979280

Soomro, Mudasar Ahmed & Memon, Rafia Naz & Chandio, Asghar Ali & Leghari, Mehwish & Soomro, Muhammad Hanif. 2024. A dataset of Roman Urdu text with spelling variations for sentence-level sentiment analysis. *Data in Brief* 57. 111170. 1–9. https://doi.org/10.1016/j.dib.2024.111170

Appendix

Writing system and transliteration

The corpus contains both Nastaliq Urdu and Roman Urdu. Nastaliq script was used for all formal sources (news and literary texts), while Roman Urdu was retained for social media and informal communication. The same scripts were preserved in the

prompts provided to GPT-4; no automatic transliteration was applied unless explicitly stated.

A.1 Corpus source overview

The dataset was compiled from publicly accessible Urdu-language sources across three domains:

(i) news media (e.g., Dawn Urdu, BBC Urdu, Express News, Jang);
(ii) social media (e.g., X (formerly Twitter) and Facebook);
(iii) open-access literary repositories and online Urdu prose collections.

Due to copyright and platform restrictions, the full dataset cannot be redistributed. However, representative examples are provided below to ensure transparency and reproducibility.

A.2 Sentence extraction and processing

Sentence extraction followed a semi-automatic procedure. Source texts were first collected at the document level, after which sentence segmentation was performed using punctuation-based and rule-based methods. The extracted sentences were then manually reviewed to ensure correctness and completeness.

"Grammatical completeness" in this study refers to sentences that contain a well-formed syntactic structure, including at minimum a subject–predicate relationship or a complete clause expressing a meaningful proposition. Fragments, headlines without predicates, and incomplete utterances were excluded.

Additional preprocessing steps included:

- removal of duplicate entries
- exclusion of non-Urdu material
- filtering of advertisements and noisy text
- normalization of Unicode characters and punctuation

Domain categorization (news, social media, literary) was carried out at the source level prior to sentence extraction.

A.3 Sample Urdu corpus

Below are representative examples of sentences included in the corpus. These examples are illustrative and reflect the diversity of domains and linguistic styles captured in the dataset.

ID	Sentence (Urdu Script)	Roman Urdu	English Translation	Domain
1	وزیراعظم نے نئی پالیسی کا اعلان کیا۔	Wazeer-e-Azam ne nai policy ka aelan kiya.	The Prime Minister announced a new policy.	News
2	آج کی کرکٹ میچ بہت دلچسپ تھی	Aaj ki cricket match bohat dilchasp thi.	Today's cricket match was very interesting.	Social media
3	زندگی ایک سفر ہے، منزلیں خود تلاش کرنی پڑتی ہیں۔	Zindagi aik safar hai, manzilen khud talash karni parti hain.	Life is a journey; one must find their own destinations.	Literary
4	صارفین نے نیا ایپلیکیشن بہت پسند کیا۔	Sarfeen ne naya application bohat pasand kiya.	Users liked the new application very much.	Technology
5	کیا آپ میرے ساتھ چلنا چاہیں گے؟	Kya aap mere sath chalna chaheinge?	Would you like to come with me?	Conversational

Note:
Short evaluative sentences (e.g., from social media) were intentionally included to capture informal and user-generated language relevant for sentiment analysis.

A.4 Example GPT-4 prompts

The following prompts illustrate the structured instruction-based approach used for semantic annotation tasks.

Entity Type PERSON: Includes personal names as well as occupational titles and institutional roles referring to individuals (e.g., Prime Minister, President, Chief Justice, Professor).

Below are sample prompts used for each semantic annotation task:

A.4.1 NER

Task: Identify all named entities in the following Urdu sentence. Classify each entity as PERSON, ORGANIZATION, LOCATION, or DATE. Provide output in JSON format.

Sentence (Urdu): وزیر اعظم نے اسلام آباد میں نئی پالیسی کا اعلان کیا۔
English translation: 'The prime minister announced a new policy in Islamabad.'
Expected output: {"entities": [{"text": "وزیر اعظم", "type": "PERSON"}, {"text": "اسلام آباد", "type": "LOCATION"}]}

A.4.2 Semantic similarity

Task: Determine the semantic similarity between the following two Urdu sentences. Provide a score between 0 (completely different) and 1 (identical meaning).

Sentence 1 (Urdu): آج موسم بہت خوشگوار ہے۔
English translation 1: 'The weather is very pleasant today'.

Sentence 2 (Urdu): موسم آج بہت اچھا ہے۔
English translation 2: 'The weather is very good today'.
Expected output: {"similarity_score": 0.92}

A.4.3 Sentiment analysis

Task: Determine the sentiment of the following Urdu text. Classify it as Positive, Negative, or Neutral.

Sentence (Urdu): مجھے یہ فلم بہت پسند آئی۔
English translation: 'I really liked this movie.'
Expected output:{"sentiment": "Positive"}

A.5 Evaluation metrics

GPT-4 Similarity scores: Scalar values between 0 and 1 generated directly by the model in response to the similarity prompt.
Reference embeddings: Dense vector representations of sentences produced using the OpenAI text-embedding-3-large model.
Cosine similarity: Used to compute vector similarity between reference embeddings, serving as a distributional semantic baseline and enabling correlation analysis with both human ratings and GPT-4 scores.

A.5.1 NER metrics

- **Precision:** Fraction of predicted entities that are correct.
- **Recall:** Fraction of true entities correctly predicted.
- **F1-score:** Harmonic mean of precision and recall.

A.5.2 Semantic similarity metrics.

- Pearson correlation: Measures correlation between GPT-4 similarity scores and human-annotated similarity ratings.
- Reference embeddings: Dense vector representations of sentences generated using the OpenAI *text-embedding-3-large* model.
- Cosine similarity: Used to compute semantic proximity between sentence embeddings in vector space. These cosine scores were correlated with both GPT-4–generated similarity scores (0–1) and human-annotated similarity ratings to assess alignment between distributional semantics, model predictions, and human judgments.

A.5.3 Sentiment analysis metrics

- **Accuracy:** Percentage of sentences correctly classified.
- **Macro-F1:** Average F1-score across all sentiment classes, accounting for class imbalance.

A.5 Sample annotated data

The following examples illustrate how annotation was applied to the dataset across different tasks.

NER Example (with English translation)

Sentence (Urdu)	English translation	Entities (Text)	Type
اوگرا نے ۔ پیٹرول کی قیمتوں میں اضافہ ہو گیا	'OGRA increased petrol prices.'	اوگرا	ORGANIZATION

Semantic Similarity example (with English Translation)

Sentence 1 (Urdu)	English translation 1	Sentence 2 (Urdu)	English translation 2	GPT-4 Similarity score	Human score
آج موسم خوشگوار ہے	'The weather is pleasant today.'	موسم آج بہت اچھا ہے۔	'The weather is very good today.'	0.92	0.95

Sentiment analysis example (with English translation)

Sentence (Urdu)	English translation	Predicted sentiment	True sentiment
یہ کتاب بہت بورنگ تھی	'This book was very boring.'	NEGATIVE	NEGATIVE
میں نے پارٹی کا بہت لطف اٹھایا۔	'I really enjoyed the party.'	POSITIVE	POSITIVE

Challenging sentiment cases

Sentence (Urdu)	English translation	Sentiment type	Challenge
فلم اچھی تھی لیکن بہت لمبی تھی	'The movie was good but very long.'	MIXED	conflicting polarity
واہ کیا شاندار سروس ہے!	'Wow, what excellent service!'	SARCASTIC	pragmatic meaning
ٹھیک تھا، لیکن زیادہ خاص نہیں	'It was okay, but nothing special'	NEUTRAL-NEGATIVE	subtle evaluation

A.7 Annotation guidelines

Annotators followed predefined guidelines specifying:

- entity categories (PERSON, ORGANIZATION, LOCATION, DATE)
- sentiment classes (Positive, Negative, Neutral)
- similarity scoring criteria (0-1 scale)

These guidelines ensured consistency across human annotations

A.8 Corpus statistics

Feature	Count/Percentage
Total sentences	50,000
News sentences	15,000
Social media sentences	20,000
Literary texts	15,000
Average words per sentence	12
Percentage of Roman Urdu	18%

A.9 Hardware and computational details

- Model Access: GPT-4 API via OpenAI platform (no fine-tuning applied).
- Batch Size: 100 sentences per API call.
- Processing Time: Approx. 1,500 sentences/hour.
- Post-processing: JSON outputs validated and normalized using Python scripts.

A.10 Limitations of appendix data

- The sample sentences are illustrative; the full corpus contains more linguistic diversity.
- Roman Urdu representation may not cover all spelling variations in informal social media text.
- GPT-4 outputs may vary slightly depending on prompt phrasing and API version.

- The examples presented in this appendix are illustrative and do not fully capture the complete diversity of the dataset. Roman Urdu representation may not include all spelling variations found in informal communication. Additionally, GPT-4 outputs may vary slightly depending on prompt phrasing and model updates.