Named Entity Recognition for Persian Literary Text: A Case Study on The Little Prince

Minoo Nassajian, Joakim Nivre, Daniel Zeman
Charles University, Faculty of Mathematics and Physics
Prague, Czech Republic
{nassajian, zeman}@ufal.mff.cuni.cz

Uppsala University, Department of Linguistics and Philology
Uppsala, Sweden
joakim.nivre@lingfil.uu.se

Abstract

Existing Persian Named Entity Recognition (NER) research has focused predominantly on news and social media domains, leaving literary texts—with their distinct linguistic characteristics—virtually unexplored. This paper addresses this gap by developing a new literary NER corpus using the Persian translation of *The Little Prince* story and evaluating existing state-of-the-art Persian NER tools on this corpus, trained exclusively on news and social media corpora. Our analysis reveals significant performance degradation on literary text, identifying systematic errors related to narrative-specific entities, metaphorical language, and discourse structures that challenge conventional NER approaches.

Keywords: Persian NER, Persian NLP, Persian computational literary

1. Introduction

Named entity recognition (NER), as a fundamental sub-task within the field of natural language processing (NLP), is concerned with the automatic identification and classification of entities in unstructured texts into predefined semantic categories including but are not limited to, the names of persons, organizations, and geographical locations (Grishman and Sundheim, 1996; Nadeau and Sekine, 2007).

As an important component in the NLP pipeline, the utility of NER extends far beyond mere identification. It serves as a foundational pipeline for a wide array of downstream applications and advanced research domains. Its contributions are integral to systems for information retrieval (Mandl and Womser-Hacker, 2005; Petkova and Croft, 2007; Guo et al., 2009), question answering (Mollá et al., 2006, 2007; Khalid et al., 2008), and machine translation (Babych and Hartley, 2003; Vu et al., 2020; Xie et al., 2022), where disambiguating entity references is essential for accuracy. Furthermore, NER provides a crucial input for more complex linguistic tasks such as relation extraction (Feldman and Rosenfeld, 2006; Gundluru et al., 2022), co-reference resolution (Zhao, 2009; Clark and Manning, 2016), and automatic text summarization (Khademi and Fakhredanesh, 2020; Berezin and Batura, 2022; Khan et al., 2024).

The majority of NER research has been conducted on a narrow range of domains—primarily news (Shabat and Omar, 2015; Al-Ash and Wibowo, 2018; Ruokolainen et al., 2020; Chavan and Patil, 2024), encyclopedic text (Balasuriya et al., 2009; Nothman et al., 2013; Li et al., 2019), and, more recently, social media (Aguilar et al., 2017; Nie et al., 2020; Yu et al., 2023) and clinical texts (Chen et al., 2015; Kundeti et al., 2016; Bose et al., 2021; Goyal and Singh, 2025)—where entity mentions are often explicit, capitalized, and anchored to real-world referents.

Unlike the clear, explicitly named entities common in journalistic or formal texts, literary entities often possess an ontological ambiguity—they may be fictional constructs rather than real-world referents (Bamman et al., 2019). Furthermore, their introduction and reference within the narrative are frequently implicit, relying on descriptive epithets, figurative language, or discourse-driven uniqueness, often without capitalization or explicit proper names (Silva and Moro, 2024). As a result, NER systems trained on surface-level regularities often fail to capture entities that are central to narrative meaning.

This paper investigates Persian NER in the literary domain through a detailed case study of *The Little Prince* story. We construct a Persian NER corpus, do pre-processing steps using a standard NLP pipeline, and annotate it according to a linguistically grounded framework that distinguishes strong named entities (e.g., proper names and rigid designators) from weak named entities (e.g., definite descriptions and discourse-unique referents) proposed by Borrega et al. (2007). We then evaluate three state-of-the-art Persian NER systems and compare their outputs to annotations produced via controlled in-context prompting of ChatGPT, which was instructed to follow the same theoretical guidelines. This work is part of a broader effort aimed at

Proceedings of The Seventh International Workshop on Designing Meaning Representations (DMR 2026) @ LREC 2026, pages 54–64

developing a Persian Uniform Meaning Representation (UMR) corpus based on the existing Persian AMR resource. Reliable identification of named entities is a necessary prerequisite for UMR annotation, as entity types contribute to argument interpretation, animacy distinctions, and discourse-level coreference. The literary NER corpus introduced in this paper therefore serves as an enabling resource supporting ongoing Persian UMR development.

The remainder of this paper is organized as follows: Section 2 reviews foundational and contemporary literature. Section 3 and 4 introduce the annotated corpus and the domain-specific annotation framework developed for this work. Section 5 explains the NER tagset used in this research. Section 6 outlines our experimental methodology. Finally, section 7 presents the empirical results and a systematic error analysis. Section 8 concludes with a summary of contributions and future research avenues.

2. Related Works

NER emerged as a core task in information extraction with early benchmarks such as MUC, ACE, and CoNLL, which were largely grounded in newswire and broadcast data. These datasets shaped both annotation guidelines and modeling assumptions, privileging capitalized proper names, geopolitical entities, and organizations, and favoring relatively flat syntactic structures. Consequently, most high-performing NER systems were trained on news-domain corpora and implicitly encode assumptions about journalistic style, entity distributions, and referential clarity.

Subsequent work has demonstrated that NER models trained on news data degrade substantially when applied to other domains, including social media, historical documents, and literary texts. Augenstein et al. (2017) showed that domain shift leads to sharp drops in precision and recall, especially for entities that are frequent but stylistically atypical. Literary texts pose additional challenges, including metaphor, personification, and character roles that function as names within the story world. In response to this gap, a few recent studies have been undertaken to expand the methodological toolkit of computational literary analysis.

van Dalen-Oskam et al. (2014) introduced the Namescape project as one of the earliest systematic efforts to apply NER to large-scale literary corpora. Motivated by the needs of literary scholarship rather than traditional information extraction, their work emphasized the importance of identifying both real-world and fictional entities in literary texts, highlighting challenges such as referential ambiguity, name variation, and the prevalence of non-canonical entities that are typically absent from newswire corpora. They described the construction of Dutch literary corpora derived from digitized novels and historical texts, alongside initial annotation guidelines tailored to literary discourse, but did not primarily focus on model evaluation or benchmark performance. Building on this foundation, de Does et al. (2017) extended the Namescape initiative by producing fully annotated gold-standard literary corpora and conducting systematic experiments with supervised NER models trained on these texts. Using Conditional Random Fields (CRF) (Wallach, 2004) and Support Vector Machine (SVM) (Hearst et al., 1998) models, they demonstrated that models trained on in-domain literary data substantially outperform systems trained on news-domain corpora, reporting large improvements in F_1 score and confirming the inadequacy of news-based NER assumptions for literary language.

LitBank corpus is another computational literary effort on NER, introduced by Bamman et al. (2019). This is an annotated corpus of 210,532 tokens drawn from 100 English literary works. The annotation follows the ACE 2005 guidelines (Consortium et al., 2005), but crucially extends them to literary-specific phenomena by annotating both named and common noun phrases, allowing nested entities, and incorporating imagined or fictional locations and characters.

Complementary to corpus creation, Dekker et al. (2019) provide an extensive evaluation of off-the-shelf NER systems on literary novels, focusing on the task of social network extraction from fiction. Their study evaluates four widely used NER systems on 40 English novels (20 classic, 20 modern), manually annotating approximately one chapter per novel for PERSON entities.

In a recent study, Silva and Moro (2024) introduce a manually annotated corpus designed specifically for NER in Portuguese literary texts. The corpus comprises 25 public-domain literary works from Brazilian and European Portuguese literature, each contributing approximately 5,000 tokens.

Prior research on Persian NER has predominantly focused on developing datasets and systems for non-literary domains such as news (Poostchi et al., 2016; Shahshahani et al., 2018), social media, and Wikipedia (Asgari-Bidhendi et al., 2021; Aghajani et al., 2021a). Consequently, the performance of existing tools—trained on these corpora—on the distinct linguistic and stylistic features of full-length literary fiction remains largely unexamined. Furthermore, the field lacks systematic annotation guidelines grounded in linguistic theory specifically tailored for the Persian literary domain. This constitutes a significant gap, as a systematic evaluation of Persian NER on narrative texts is essential for advancing computational literary studies and digital humanities in Persian.

The central contribution of the present work is the creation of the first manually annotated Persian literary NER corpus, which is released as a publicly available resource to support future research. The corpus is based on the Persian translation of *The Little Prince* story and it complements ongoing work on Persian AMR-to-UMR conversion by providing consistent entity annotations required for semantic role interpretation and discourse tracking. Using this gold-standard dataset, we further evaluate 3 existing state-of-the-art Persian NER systems (ParsNER (Team, 2021), Shekar (Amirivojdan, 2025), and ParsTwiNER (Aghajani et al., 2021b)) and analyze their performance in the literary domain. In the next sections, we will explain about the statistical information of the corpus and NER tagsets used in this research.

3. Corpus

Our research employs the Persian Abstract Meaning Representation (PAMR) corpus, developed by Takhshid et al. (2024). This corpus is based on the Persian translation of *The Little Prince* and it contains no prior named entity annotations. The corpus consists of 1,562 sentences (14,427 tokens) with a vocabulary of 3,520 unique word forms; sentence lengths range from 1 to 65 tokens. We preprocessed the text by normalizing and tokenizing it using the Stanza toolkit (Qi et al., 2020). The corpus was then formatted according to the BIO tagging scheme to facilitate subsequent manual named entity annotation and sequence-labeling experiments.

4. Annotation Guideline

To establish annotation guidelines suited to literary texts, we ground our approach in the theoretical framework of Borrega et al. (2007), which argues that named entities should not be defined solely by formal properties (e.g., capitalization[1]), but rather by referential behavior in discourse. Central to this framework is the distinction between Strong Named Entities (SNEs) and Weak Named Entities (WNEs), and the notion of Trigger Words (TWs), which play a decisive role in identifying entities that lack canonical proper names. In the following sections, we will discuss how we annotate *The Little Prince* story using this distinction.

4.1. Strong Named Entities

SNEs are expressions whose referent can be identified independently of discourse context, typically through a conventional name. In *The Little Prince* story, there are some words that are considered as SNEs. For example, the word "Earth" does not refer to an abstract concept or a generic surface in this text; instead, it denotes the planet Earth as a unique astronomical entity. Similarly, "Africa" refers to a specific continent with a fixed real-world referent.

4.2. Weak Named Entities

WNEs are nominal expressions that function as unique referents within a specific discourse due to contextual anchoring, despite lacking external uniqueness. Their detection is operationalized via Trigger Words (TWs). TWs are common nouns that, as semantic heads of noun phrases, signal potential entities. TWs require discourse-specific modifiers—such as definiteness, relational adjectives, or narrative uniqueness—to trigger entity interpretation. A noun phrase is annotated as a WNE if it meets these criteria:

- Head: Contains a Trigger Word
- Uniqueness: Referent is unique in the discourse
- Trackability: Referent is maintained across mentions
- Role: Referent holds a distinct narrative or semantic role

A clear example of a WEN is found in example (1). The common noun ("flower") functions as the Trigger Word. Through the discourse, this nominal expression becomes a unique and trackable entity. This occurs as the definite reference ("the flower") first establishes specificity, after which subsequent mentions ("that flower") maintain a coherent coreferential chain. Ultimately, by functioning as a central character within the narrative, it gains significant semantic weight. Thus, despite its common noun head, the phrase satisfies all criteria for annotation as a WEN.

(1) "I am not a weed," the flower replied.[2]

This annotation principle applies consistently to other WNEs in *The Little Prince*, including "the king", "the fox", and "the geographer". Each of these nominal expressions, headed by a common noun Trigger Word, similarly acquires unique referential status through definiteness, narrative anchoring, and sustained discourse presence.

[1] Unlike Latin-script languages, Persian orthography does not employ capitalization to distinguish proper nouns.

[2] In this work, WNEs are limited to noun phrases. Pronouns were not annotated as WNEs, even when they refer to uniquely identifiable entities, as they function primarily as coreferential expressions rather than entity-denoting mentions.

5. NER tagset

The named entity inventory employed in this study encompasses a deliberately broad range of semantic categories, some of which are quite specific: **Person, Animal, Plant, Organization, Location, Product, Publication, Nationality, Time, and Planet.**[3] Following the standard BIO tagging convention, tokens that do not belong to any named entity span were assigned the label **O**, which represents non-named-entity tokens. Table 1 shows the entity distribution in the Persian corpus of *The Little Prince*.

Entity Type	Entities
Person-W	261
Animal-W	43
Plant-W	45
Planet	33
Location	24
Time	17
Product	8
Nationality	3
Publication	2
Organization	1
Total	**437**

Table 1: Entity Distribution in persian corpus of *The Little Prince*

5.1. Inter-Annotator Agreement

To further assess the reliability of the gold annotations, a second linguist independently annotated a subset of 100 sentences selected to include diverse named entity types. Inter-annotator agreement was measured at the token level using Cohen's κ coefficient. The resulting agreement score was $\kappa = 0.99$, indicating almost perfect agreement. Disagreements were resolved through adjudication with a third annotator, and the finalized labels were incorporated into the gold dataset.

6. Experimental Setup

In this work, we evaluate the robustness of existing Persian NER tools on the preprocessed Persian corpus of *The Little Prince*. In the next parts, we will explain the pre-processing steps and Persian NER tools.

[3]Within the annotation schema, the SNE/WNE distinction is implemented via a dedicated suffix. The tag -W is appended to standard BIO tags to flag a Weak Named Entity (e.g., B-person vs. B-person-W)

6.1. Preprocessing and Input Standardization

The Persian Little Prince corpus was preprocessed using the Persian pipeline in Stanza, including text normalization and tokenization. The processed corpus was then converted into BIO sequence-labeling format for named entity annotation and evaluation. Since our goal is to assess the behavior of existing Persian NER taggers on literary text, we treat the entire corpus as test-only data and report model performance over all sentences.

6.2. Automatic Annotation Systems

To assess how existing Persian NER systems perform on literary narrative text, we evaluate three state-of-the-art transformer-based Persian NER tools—ParsNER (Asgari-Bidhendi et al., 2021), Shekar (Amirivojdan, 2025), and ParsTwiNER (Aghajani et al., 2021b)—all of which were originally developed and trained on non-literary corpora such as news, Wikipedia, or social media. These systems differ substantially in model architecture, training data, and tagset granularity, reflecting their original design goals and target domains.

In addition, we evaluate a guideline-driven LLM-based annotator (ChatGPT-5.2, OpenAI), prompted to follow a linguistically motivated NER framework distinguishing SNEs and WNEs, along with a fixed set of 20 annotated examples drawn from different chapters of the text. These examples were selected to cover diverse entity types and narrative phenomena typical of literary discourse. Moreover, These sentences were excluded from the evaluation set to prevent prompt–test overlap and ensure unbiased performance measurement. Remaining unseen portion of the corpus were automatically annotated and subsequently reviewed by authors, who manually corrected incorrect labels and added missing entity annotations to address both precision and recall errors. This process ensured consistency with the annotation guidelines and resolved systematic errors, particularly in the identification of weak named entities and discourse-dependent references. The resulting annotations constitute a manually validated gold dataset.

6.3. Label Mapping for Cross-System Comparability

Because the compared systems differ substantially in tagset and granularity, we conduct evaluation using two complementary protocols. In the primary evaluation ("shared tagset"), all outputs were mapped to a common label inventory consisting of

PER, LOC,[4] ORG, DAT, and O. Dataset-specific labels not belonging to this inventory (e.g., MON, PCT, PRO; POG; or fine-grained narrative categories) were mapped to O.

In addition to the shared-tagset comparison, we perform a second evaluation protocol focusing on ChatGPT's guideline-driven annotation scheme. Unlike the automatic taggers, ChatGPT was instructed to annotate literary entities using a fine-grained label set (introduced in NER tagset section) that distinguishes both entity category and referential status (Strong vs. Weak NEs).

6.4. Metrics

We report entity-level Precision (P), Recall (R), and F_1 using BIO span evaluation as implemented in the seqeval library. Micro-averaged F_1 over entity types is used as the main metric due to label imbalance and the predominance of O tokens, while macro-averaged scores and per-class F_1 are also reported for interpretability. Confusion matrices (row-normalized) are provided to characterize systematic confusions as well.

7. Results and Error Analysis

Table 2 presents entity-level Precision, Recall, and F_1-score for the three Persian NER taggers (ParsNER, ParsTwiNER, and Shekar) and the LLM-based guideline-driven annotation system (ChatGPT) on the Persian Little Prince corpus. All outputs were normalized into a unified tagset (PER/LOC/ORG/DAT), and evaluation was performed using BIO entity spans.

Across the automatic systems, ParsNER achieves the strongest performance among the supervised taggers (micro-$F_1 = 0.52$), while ParsTwiNER and Shekar degrade sharply (micro-$F_1 = 0.21$ and 0.09), indicating severe domain shift from their original training corpora (news/Wikipedia/Twitter) to literary narrative. In contrast, ChatGPT substantially outperforms all supervised taggers, reaching micro-$F_1 = 0.90$, driven primarily by very high recall (0.92). This confirms that a guideline-constrained LLM annotation protocol is particularly suitable in literary settings, and NER models trained on news, Wikipedia, or Twitter do not transfer reliably to Persian literary narrative, where entity mentions are often discourse-dependent, stylistically variable, and frequently realized as non-canonical noun phrases.

The per-class analyses, presented in Tables 3, 4, 5, and 6, illustrate the nature of domain mismatch. ParsNER achieves high precision for LOC (0.93) but only moderate recall (0.39), indicating that it detects a subset of easily recognizable locations but fails on narrative or ambiguous location mentions. ParsTwiNER, optimized for social media, demonstrates catastrophic performance on literary character recognition (PER recall $= 0.08$, $F_1 = 0.13$) and completely fails to identify temporal expressions (DAT recall $= 0.00$). While it achieves reasonable precision for locations (0.85), its recall remains low (0.34), suggesting it recognizes only the most conventional location mentions. Shekar, trained on encyclopedic Wikipedia text, shows a different pathology: it achieves moderate recall for dates (0.41) but with near-zero precision (0.05), indicating severe over-prediction of temporal expressions. Most critically, it fails entirely to recognize character mentions (PER recall $= 0.00$, $F_1 = 0.01$), rendering it fundamentally unsuitable for narrative analysis.

System	P	R	Micro-F1	Macro-F1
ParsNER	0.64	0.43	0.52	0.30
ParsTwiNER	0.48	0.13	0.21	0.32
Shekar	0.12	0.07	0.09	0.22
ChatGPT	**0.88**	**0.92**	**0.90**	**0.55**

Table 2: Entity-level performances of Persian NER systems on the *Little Prince* corpus

Class	P	R	F1	Support
DAT	0.20	0.06	0.09	15
LOC	0.93	0.39	0.55	55
ORG	0.20	1.00	0.33	1
PER	0.63	0.46	0.53	255

Table 3: Class-wise performance metrics for the ParsNER system on the *Little Prince* corpus.

In contrast, Table 6 shows that ChatGPT exhibits particularly high effectiveness on the dominant entity categories in the narrative, reaching F_1 scores of 0.97 for PER and 0.82 for LOC, indicating that it successfully detects nearly all character and location mentions. This supports the hypothesis that large language models can generalize entity recognition beyond the news domain and remain highly sensitive to narrative discourse structure. In contrast, DAT is substantially harder ($F_1 = 0.40$), reflecting the ambiguity of temporal expressions in Persian literary text, where time is frequently encoded through discourse-relative adverbs (e.g., tonight, tomorrow, that night) rather than explicit calendar dates, and the low macro-F_1 (0.55) is largely driven by these low-frequency and ambiguous classes. These findings suggest that guideline-driven ChatGPT annotation is particularly suitable for corpus bootstrapping in literary settings, where maximizing entity coverage (recall) is critical for constructing reliable gold datasets.

Based on Figures 1 and 2, we can see a clear performance gap between the two systems, especially

[4] References to celestial bodies (planets) and nationalities within *The Little Prince* were aligned with the LOC label.

Class	P	R	F1	Support
DAT	0.00	0.00	0.00	15
LOC	0.85	0.34	0.49	55
ORG	0.50	1.00	0.67	1
PER	0.33	0.08	0.13	255

Table 4: Class-wise performance metrics for the ParsTwiNER system on the *Little Prince* corpus.

Class	P	R	F1	Support
DAT	0.05	0.41	0.09	15
LOC	0.35	0.24	0.28	55
ORG	0.33	1.00	0.50	1
PER	0.12	0.00	0.01	255

Table 5: Class-wise performance metrics for the Shekar system on the *Little Prince* corpus.

for PER and DAT entities. ChatGPT demonstrates near-ceiling recognition of PER mentions (98.0%), while ParsNER correctly predicts only 57.2% of PER tokens and incorrectly maps a substantial portion to the non-entity class (O; 42.5%). This indicates that ParsNER systematically fails to recover character mentions in narrative discourse, due to genre mismatch and its reliance on news-style naming conventions. For LOC, both models show confusion with O, but ChatGPT achieves substantially higher accuracy (81.5%) compared to ParsNER (42.0%), suggesting stronger contextual inference for spatial references in descriptive passages. Finally, both models show difficulty with temporal expressions (DAT), though ChatGPT retains moderate accuracy (62.9%) whereas ParsNER almost entirely collapses DAT into O (94.3%). Overall, the confusion matrix comparison confirms that ChatGPT is substantially more robust to literary-domain entity realizations, while ParsNER exhibits heavy domain-shift degradation, especially for discourse-driven person references and non-canonical time mentions.

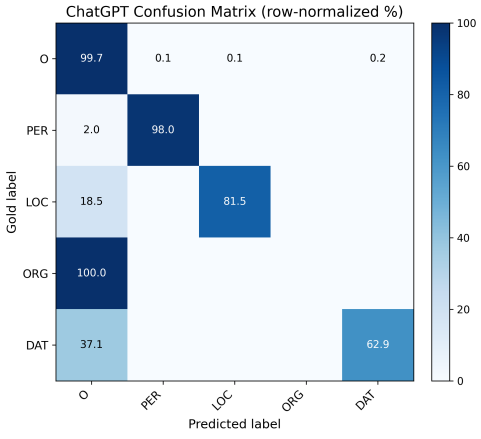

Figure 1: ChatGPT Confusion Matrix (Mapped Tagset)

Class	P	R	F1	Support
DAT	0.30	0.59	0.40	15
LOC	0.82	0.82	0.82	55
ORG	0.00	0.00	0.00	1
PER	0.97	0.97	0.97	255

Table 6: Class-wise performance metrics for the ChatGPT system on the *Little Prince* corpus.

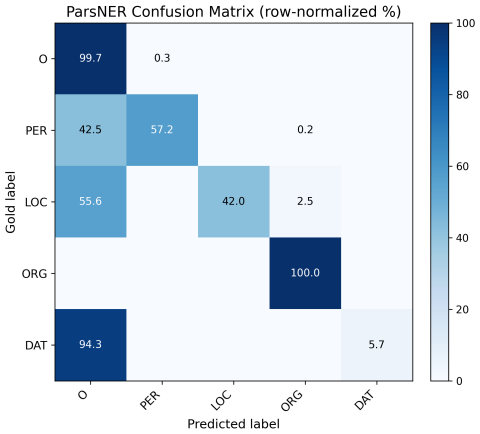

Figure 2: ParsNER Confusion Matrix (Mapped Tagset)

Under Protocol 2, which evaluates performance on the full fine-grained tagset without label mapping, ChatGPT demonstrates exceptionally strong performance on the Persian Little Prince corpus, achieving a micro-averaged F_1-score of 0.87. Table 7 reveals the model's particular proficiency with narrative-specific, discourse-grounded entities. Most notably, it achieves near-perfect recognition of weak person references (person-W) with an F_1-score of 0.97 on substantial support ($n = 255$). This indicates a robust capacity to track characters through implicit, descriptive, and role-based mentions—a core challenge in literary analysis. The model also excels at identifying other distinctive narrative entities, such as speaking animals (animal-W $F_1 = 0.92$) and symbolic flora (plant-W $F_1 = 0.79$), alongside strong performance for conventional categories like location ($F_1 = 0.82$). However, performance remains uneven across the tagset.

Temporal expressions (time $F_1 = 0.40$) and several rare categories with minimal support—such as organization, product, and nationality—yield low or unstable scores. The pronounced discrepancy between the macro-averaged (0.53) and micro-averaged (0.87) F_1-scores underscores this class-wise variance, reflecting the challenge of generalizing across infrequent entity types. Moreover, plant has support $n = 0$, meaning that this label does not occur in the gold annotations; consequently, it cannot receive a meaningful preci-

sion/recall score. These results demonstrate that a guideline-driven LLM is a highly capable annotator for literary NER. Its strength lies in accurately recognizing the discourse-anchored entities—such as characters referred to by description rather than name—that are essential for understanding narrative but are poorly captured by conventional models.

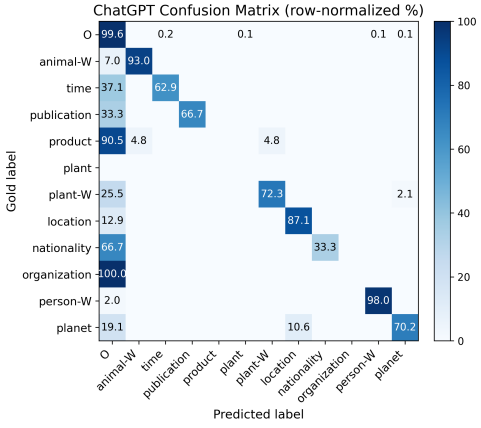

Figure 3: ChatGPT Confusion Matrix (Full Tagset)

Figure 3 shows the row-normalized confusion matrix for ChatGPT under Protocol 2 (full tagset). The model exhibits its strongest separability on discourse-central weak entities, especially person-W (98% correctly predicted) and animal-W (93%), confirming reliable tracking of recurring narrative referents. Performance is also strong for location (87%), while planet is only moderately distinguished (70%), with frequent confusion into O (e.g. the sun[5]) and occasionally location (e.g. the Earth). The main weaknesses occur in categories that are either rare in the corpus or linguistically ambiguous, such as time (substantial confusion into O due to relative temporal expressions), and sparsely represented classes like organization, nationality, and product, which are often collapsed into O[6] Overall, the matrix indicates that ChatGPT is particularly effective for highly salient discourse entities (weak NEs) but less stable on low-frequency or boundary-vague categories.

[5]The *planet* category is used as a general label for celestial bodies mentioned in the narrative, including planets and stars such as Earth and the Sun. This operational definition prioritizes annotation consistency and referential function over strict astronomical classification.

[6]Some entity types appear very rarely or not at all in the gold annotations (for example, the *plant* category has a support of 0), which makes their performance less reliable and explains the unstable patterns observed in the confusion matrix (see Table 7 for support values).

8. Conclusion and Future Work

This paper presented a corpus-based study of Named Entity Recognition in Persian literary text and introduces the first linguistically validated annotated corpus derived from *The Little Prince*, which is made freely available for research.[7] High inter-annotator agreement demonstrates the reliability of the proposed annotation scheme and establishes a solid foundation for future extensions of literary NER resources in Persian.

We also developed an annotation framework grounded in the theoretical distinction between strong and weak named entities, operationalized through a discourse-sensitive tagset designed to capture referential patterns characteristic of narrative texts. Using the gold-standard corpus, we conducted an empirical evaluation comparing three state-of-the-art Persian NER systems with a guideline-driven annotation approach implemented using ChatGPT.

The experimental results reveal a pronounced domain shift challenge: existing pretrained taggers exhibited significant performance degradation when applied to literary text, attributable to mismatches in referential structure, entity realization, and Persian-specific linguistic ambiguity. In contrast, the LLM-based approach achieved substantially higher overall F_1-scores. This suggests that large language models, when guided by formal annotation criteria and supported by human validation, can serve as an effective tool for both corpus construction and the study of discourse-mediated namedness.

Future work may enrich the corpus with complementary linguistic layers, such as coreference chains, nested entity structures, and quotation attribution. These additions, which interact fundamentally with narrative entities, would enable more comprehensive modeling of Persian literary discourse.

9. Ethics Statement

We are not aware of any ethical concerns related to this work. The corpus was manually annotated as part of academic research. In addition to the primary annotator, independent linguistically trained annotators contributed to annotation validation and agreement assessment. All annotation work was conducted voluntarily for research purposes, and no sensitive or personal data were involved, as the corpus is based on a publicly available literary text.

10. Limitations

Several limitations should be considered when interpreting the findings. First, the corpus is based on

[7]http://hdl.handle.net/11234/1-6136

Entity Type	Precision	Recall	F1	Support
animal-W	0.91	0.93	0.92	40
location	0.75	0.91	0.82	21
nationality	1.00	0.33	0.50	3
organization	0.00	0.00	0.00	1
person-W	0.97	0.97	0.97	255
planet	0.76	0.71	0.73	31
plant	0.00	0.00	0.00	0
plant-W	0.89	0.71	0.79	43
product	0.00	0.00	0.00	6
publication	1.00	0.50	0.67	2
time	0.30	0.59	0.40	15
Micro Avg	**0.86**	**0.87**	**0.87**	**417**
Macro Avg	**0.60**	**0.51**	**0.53**	**417**
Weighted Avg	**0.88**	**0.87**	**0.87**	**417**

Table 7: ChatGPT performance on the Persian *Little Prince* corpus under Protocol 2 (full tagset evaluation without label mapping). "-W" denotes weak named entities (WNEs). Note that the *plant* category has a support of 0 in the gold annotations.

a single literary work, which limits stylistic diversity and may not fully represent entity behavior across Persian literature. While *The Little Prince* contains a rich set of narrative referents, it is structurally simpler than many Persian novels and may underrepresent complex constructions such as heavy embedding, poetic metaphor, and culturally grounded naming patterns. Future expansions to additional works are therefore necessary for broader generalization.

Second, entity label distribution is imbalanced. PER and LOC dominate the annotated mentions, whereas ORG occurs rarely. As a result, performance metrics for low-frequency labels are unstable, and confusion matrix interpretation is more reliable for frequent categories. This imbalance also impacts macro-averaged scores, which penalize models heavily for rare labels even when those labels contribute little to overall entity mass.

Finally, evaluation requires label mapping across heterogeneous tagsets. Shared-tag evaluation provides a fair comparison but necessarily collapses distinctions in the original tools (e.g., ParsNER's extended categories such as money/percent/product). Conversely, ChatGPT's richer tagset (including animal, plant, planet) cannot be directly compared under standard NER metrics without multi-label mapping. Therefore, our results should be interpreted as two complementary analyses: (i) standardized shared-tag performance evaluation and (ii) descriptive coverage statistics for the full tagsets.

Acknowledgments

The work described herein was supported by the Charles University, project GAUK No. 394625. The second author was also funded by *LINDAT/CLARIAH-CZ* (Project No. LM2023062) of the Ministry of Education, Youth, and Sports of the Czech Republic. This research was also partially supported by SVV project number 260 821.

11. Bibliographical References

MohammadMahdi Aghajani, AliAkbar Badri, and Hamid Beigy. 2021a. Parstwiner: A corpus for named entity recognition at informal persian. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 131–136.

MohammadMahdi Aghajani, AliAkbar Badri, and Hamid Beigy. 2021b. ParsTwiNER: A corpus for named entity recognition at informal Persian. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 131–136, Online. Association for Computational Linguistics.

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153.

Herley Shaori Al-Ash and Wahyu Catur Wibowo. 2018. Fake news identification characteristics using named entity recognition and phrase detection. In *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 12–17. IEEE.

Ahmad Amirivojdan. 2025. Shekar: A Python Toolkit for Persian Natural Language Processing. *Journal of Open Source Software*, 10(114):9128.

Majid Asgari-Bidhendi, Behrooz Janfada, OR Roshani Talab, and Behrouz Minaei-

Bidgoli. 2021. Parsner-social: A corpus for named entity recognition in persian social media texts. *Journal of AI and Data Mining*, 9(2):181–192.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R Curran. 2009. Named entity recognition in wikipedia. In *Proceedings of the 2009 workshop on the people's web meets NLP: Collaboratively constructed semantic resources (People's Web)*, pages 10–18.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144.

Sergey Berezin and Tatiana Batura. 2022. Named entity inclusion in abstractive text summarization. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 158–162.

Oriol Borrega, Mariona Taulé, and M Antònia Martí. 2007. What do we mean when we speak about named entities. In *Proceedings of Corpus Linguistics*, pages 1–27. Citeseer.

Priyankar Bose, Sriram Srinivasan, William C Sleeman IV, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18):8319.

Tejal Chavan and Seema Patil. 2024. Named entity recognition (ner) for news articles. *Dev.(IJAIRD)*, 2(1):103–112.

Yukun Chen, Thomas A Lasko, Qiaozhu Mei, Joshua C Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18.

Kevin Clark and Christopher D Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. *arXiv preprint arXiv:1606.01323*.

Linguistic Data Consortium et al. 2005. Ace (automatic content extraction) english annotation guidelines for entities. *Version*, 5(6):2005–08.

Jesse de Does, Katrien Depuydt, Karina Van Dalen-Oskam, Maarten Marx, et al. 2017. Namescape: named entity recognition from a literary perspective. *CLARIN in the Low Countries*, pages 361–370.

Niels Dekker, Tobias Kuhn, and Marieke van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, 5:e189.

Ronen Feldman and Benjamin Rosenfeld. 2006. Boosting unsupervised relation extraction by using ner. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 473–481.

Nandita Goyal and Navdeep Singh. 2025. Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions. *Neurocomputing*, 618:129171.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Nagaraja Gundluru, Dharmendra Singh Rajput, Kuruva Lakshmanna, Rajesh Kaluri, Mohammad Shorfuzzaman, Mueen Uddin, and Mohammad Arifin Rahman Khan. 2022. Enhancement of detection of diabetic retinopathy using harris hawks optimization with deep learning model. *Computational Intelligence and Neuroscience*, 2022(1):8512469.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. Named entity recognition in query. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.

Mohammad Ebrahim Khademi and Mohammad Fakhredanesh. 2020. Persian automatic text summarization based on named entity recognition. *Iranian Journal of Science and Technology,*

Transactions of Electrical Engineering, pages 1–12.

Mahboob Alam Khalid, Valentin Jijkoun, and Maarten De Rijke. 2008. The impact of named entity normalization on information retrieval for question answering. In *European Conference on Information Retrieval*, pages 705–710. Springer.

Imaad Zaffar Khan, Amaan Aijaz Sheikh, and Utkarsh Sinha. 2024. Graph neural network and ner-based text summarization. *arXiv preprint arXiv:2402.05126*.

Srinivasa Rao Kundeti, J Vijayananda, Srikanth Mujjiga, and M Kalyan. 2016. Clinical named entity recognition: Challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1937–1945. IEEE.

Maolong Li, Qiang Yang, Fuzhen He, Zhixu Li, Pengpeng Zhao, Lei Zhao, and Zhigang Chen. 2019. An unsupervised learning approach for ner based on online encyclopedia. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*, pages 329–344. Springer.

Thomas Mandl and Christa Womser-Hacker. 2005. The effect of named entities on effectiveness in cross-language information retrieval evaluation. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 1059–1064.

Diego Mollá, Menno Van Zaanen, and Steve Cassidy. 2007. Named entity recognition in question answering of speech data. In *Proceedings of the 2007 Australasian Language Technology Workshop*, pages 57–65. ALTA.

Diego Mollá, Menno Van Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *Proceedings of the Australasian language technology workshop 2006*, pages 51–58.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. *arXiv preprint arXiv:2010.15458*.

Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R Curran. 2013. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175.

Desislava Petkova and W Bruce Croft. 2007. Proximity-based document representation for named entity retrieval. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 731–740.

Hanieh Poostchi, Ehsan Zare Borzeshi, Mohammad Abdous, and Massimo Piccardi. 2016. Personer: Persian named-entity recognition. In *COLING 2016-26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2020. A finnish news corpus for named entity recognition. *Language Resources and Evaluation*, 54(1):247–272.

Hafedh Ali Shabat and Nazlia Omar. 2015. Named entity recognition in crime news documents using classifiers combination. *Middle-East Journal of Scientific Research*, 23(6):1215–1221.

Mahsa Sadat Shahshahani, Mahdi Mohseni, Azadeh Shakery, and Heshaam Faili. 2018. Peyma: A tagged corpus for persian named entities. *arXiv preprint arXiv:1801.09936*.

Mariana O Silva and Mirella M Moro. 2024. Pportal_ner: An annotated corpus of portuguese literary entities. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12927–12937.

Reza Takhshid, Tara Azin, Razieh Shojaei, and Mohammad Bahrani. 2024. Persian Abstract Meaning Representation: Annotation guidelines and gold standard dataset. In *Proceedings of the 2024 UMR Parsing Workshop*, pages 8–15, Boulder, Colorado. Association for Computational Linguistics.

Hooshvare Team. 2021. Pre-trained ner models for persian. https://github.com/hooshvare/parsner.

Karina van Dalen-Oskam, Jesse de Does, Maarten Marx, Isaac Sijaranamual, Katrien Depuydt, Boukje Verheij, and Valentijn Geirnaert. 2014. Named entity recognition and resolution for literary studies. *Computational Linguistics in the Netherlands Journal*, 4:121–136.

Van-Hai Vu, Quang-Phuoc Nguyen, Kiem-Hieu Nguyen, Joon-Choul Shin, and Cheol-Young Ock.

2020. Korean-vietnamese neural machine translation with named entity recognition and part-of-speech tags. *IEICE TRANSACTIONS on Information and Systems*, 103(4):866–873.

Hanna M Wallach. 2004. Conditional random fields: An introduction.

Shufang Xie, Yingce Xia, Lijun Wu, Yiqing Huang, Yang Fan, and Tao Qin. 2022. End-to-end entity-aware neural machine translation. *Machine Learning*, 111(3):1181–1203.

Jianfei Yu, Ziyan Li, Jieming Wang, and Rui Xia. 2023. Grounded multimodal named entity recognition on social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9141–9154.

Jun Zhao. 2009. A survey on named entity recognition, disambiguation and cross-lingual coreference resolution. *Journal of Chinese Information Processing*, 23(2):3–17.