Evaluating Large Language Models for Character Identification in Italian Renaissance Epics

A Case Study on "Orlando Furioso"

Gaudenzia Genoni[1], Alfio Ferrara[1,2] and Stefano Montanelli[1]

[1]University of Milan, Department of Computer Science, Via Celoria, 18-20133 Milan, Italy
[2]University of Milan, Department of Literary Studies, Philology and Linguistics, Via Festa del Perdono, 7-20122 Milan, Italy

Abstract

In this exploratory work, we evaluate the applicability of Large Language Models to character identification in the low-resource narrative domain of Italian Renaissance epic poetry, using Ludovico Ariosto's *Orlando Furioso* as a case study. To mitigate contextual fragmentation, our pipeline integrates recursive narrative memory generation and character extraction. Experimental results are promising (F1 score of 0.758 with `Qwen-2.5-72B`), but qualitative error analysis reveals that performance remains constrained, in particular, by cataphoric resolution failures where character identities are disclosed non-linearly. We conclude by reflecting on new research directions in light of these findings.

Keywords

Narrative Understanding, Large Language Models, Character Identification, Italian Renaissance Epics

1. Introduction

The identification of named entities is a fundamental task in computational narrative understanding [1, 2]. Traditionally, it has been addressed through approaches ranging from semi-automatic methods [3] to deep learning techniques [4], with Large Language Models increasingly adopted in recent years [5]. Despite this growing body of work, little attention has been paid to the suitability of LLMs for character identification within the narrative genre of epic poetry. In particular, while computational analysis has been applied to epic narratives such as the Indian *Mahābhārata* [6], Italian Renaissance epics remain unexplored.

From a linguistic standpoint, 16th-century Italian chivalric epic poems in *ottava rima* represent a low-resource domain that poses challenges for contemporary LLMs. Previous work [7] has shown that LLMs struggle with Named Entity Recognition in 19th-century scholarly Italian prose. In poetic texts, these same difficulties, arising from archaic vocabulary and diachronic semantic shifts, are further compounded by syntactic distortions imposed by rhyme, meter, and literary conventions [8, 9]. Beyond linguistic complexity, character identification in epic poetry is constrained by LLM context window limits, necessitating segmentation into smaller textual units. This fragmentation raises obstacles for character coreference resolution [10, 11], here regarded as integral to character identification alongside mention extraction. The current state-of-the-art cross-context coreference resolution system, xCoRe [12], is not directly transferable to our setting, as it is trained on modern English prose [13, 14] and does not consider historical Italian language.

In this exploratory work, we provide an evaluation of the applicability and limitations of LLM-based character identification within Italian Renaissance epics, enriching the discussion on narrative understanding at the level of fine-grained narrative structure. We model our approach on the needs of literary scholars and adopt Ludovico Ariosto's *Orlando Furioso* as a case study, following our earlier work on the text [15]. In particular, we contribute: a) an original annotation of the poem's first twelve cantos,

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'26 Workshop, Delft (The Netherlands), 29-March-2026

gaudenzia.genoni@unimi.it (G. Genoni); alfio.ferrara@unimi.it (A. Ferrara); stefano.montanelli@unimi.it (S. Montanelli)
0009-0003-1561-1330 (G. Genoni); 0000-0002-4991-4984 (A. Ferrara); 0000-0002-6594-6644 (S. Montanelli)

to remedy the lack of specialized datasets (§ 2); b) a character identification pipeline that integrates recursive narrative memory to mitigate contextual fragmentation (§ 3); and c) a detailed error analysis of the results to guide future research (§ 5). Our code, together with the ground truth dataset, the prompts used, and supplementary material, is made available to the research community[1].

2. Data Annotation

To our knowledge, no existing datasets are available for computational studies on character extraction in Italian Renaissance epics. We therefore constructed a **ground-truth dataset** covering the first twelve cantos of *Orlando Furioso*, based on the digitized edition of the poem [16], for a total of 982 stanzas (7,856 lines)[2]. The annotation process was carried out by one expert in the literary domain. According to our annotation protocol, the annotator recorded all character mentions appearing as (a) proper names (onomastic mentions), (b) nouns/noun phrases, and (c) pronouns (including enclitic and reflexive forms). To retain only information relevant to literary analysis, we further delimited the scope of character identification according to four constraints: (1) *narratological level*, excluding all extradiegetic figures (such as the narrator and dedicatee, who function outside the primary story world [17]); (2) *individual identity*, excluding collective groups; (3) *proper-name attribution*, excluding nameless minor characters; and (4) *entity type*, excluding non-human agents. Every identified mention was resolved to the character's canonical proper name, and multiple mentions of the same character within a stanza were collapsed into a single entity instance. For clarity, annotation examples are provided in Table 1. In total, the dataset consists of 64 unique characters and 1,890 character occurrences; 106 stanzas were found to contain no explicit character mentions.

3. Methodology

We structure the **character identification pipeline** (Figure 1) around two tasks: (1) narrative memory generation and (2) character extraction. Inputs are batches of n stanzas derived from partitioning the poem's cantos (where n must be optimized experimentally: see § 4).

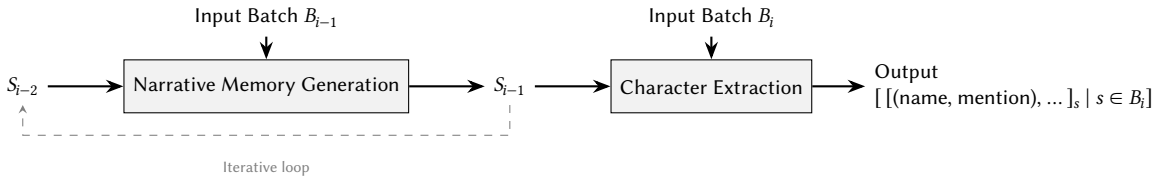

Figure 1: Overview of the character identification pipeline processing a generic batch B_i. *Narrative Memory Generation* is shown for iteration $i-1$, producing summary S_{i-1} from batch B_{i-1} and the previous summary S_{i-2}. *Character Extraction* utilizes the generated summary S_{i-1} to extract and resolve character mentions in the current batch B_i. The final output consists of pairs of proper names and textual mentions for each character retrieved within every stanza s of the batch.

1) Narrative memory generation is a recursive process. At each iteration i, an LLM produces a 2–3 sentence summary S_i based on the current batch B_i and the previous summary S_{i-1}. The resulting summary S_i serves two purposes: it provides context for the next iteration of memory generation, and it supports character identification in the subsequent batch B_{i+1}[3].

2) Character extraction is performed by a separate LLM. For each batch B_i, the model processes the text alongside the previous batch's summary S_{i-1}[4]. For each stanza within the batch, the LLM must

[1] https://github.com/umilISLab/ISLab-Text2Story26.git

[2] The first twelve cantos actually comprise 1,036 stanzas, but we omitted (from both annotation and evaluation) stanzas 22–62 of Canto III and 77–89 of Canto X: the inclusion of these sections, which consist of a genealogical catalogue and a heraldic enumeration, would have introduced a surplus of peripheral characters, overly burdening the annotation process.

[3] Although existing synopses of *Orlando Furioso* could be used, we rely on LLM-generated summaries to maintain an unsupervised setting that generalizes to lesser-studied works for which such resources are unavailable.

[4] For the first batch of Canto I (B_1), the memory S_0 is initialized with the anchor phrase "*Inizio del poema*" ("The poem begins").

Table 1
Examples of data annotation for stanzas I.3, I.5, I.12, and I.48 from Ludovico Ariosto's *Orlando Furioso* [16]. Sir Harington's English translation [18] is provided for non-Italian readers. The **Characters** column lists the character names recorded in our ground truth for each stanza. Mentions are color-coded by character identity, with subscripts marking grammatical categories: *proper name* (PN), *noun* (N), *noun phrase* (NP), and *pronoun* (P). Mentions in gray represent entities excluded per our constraints: narrator/dedicatee (I.3), collective groups (I.5), a non-human agent (I.12), an unnamed minor character (I.48).

Stanza	English Translation	Characters
Piacciavi, generosa Erculea prole, ornamento e splendor del secol nostro, Ippolito, aggradir questo che vuole e darvi sol può l'umil servo vostro. Quel ch'io vi debbo, posso di parole pagare in parte, e d'opera d'inchiostro; né che poco io vi dia da imputar sono; che quanto io posso dar, tutto vi dono.	Vouchsafe (O Prince of most renowmed race, The ornament and hope of this our time) T'accept this gift presented to your grace, By me your servant rudely here in rime. And though I paper pay and inke, in place Of deeper debt, yet take it for no crime: It may suffise a poore and humble debter, To lay and if he could it should be better.	-
Orlando$_{PN}$, che gran tempo inamorato fu de la bella Angelica$_{PN}$, e per lei$_{P}$ in India, in Media, in Tartaria lasciato avea infiniti et immortal trofei, in Ponente con essa$_{P}$ era tornato, dove sotto i gran monti Pirenei con la gente di Francia e de Lamagna re Carlo$_{NP}$ era attendato alla campagna,	Orlando$_{PN}$ who long time had loved deare, Angelica$_{PN}$ the faire: and for her sake, About the world, in nations far and neare, Did high attempts performe and undertake, Returnd with her$_{P}$ into the West that yeare, That Charles$_{PN}$ his power against the Turks did make: And with the force of Germanie and France, Neare Pyron Alpes his standard did advance.	ANGELICA, CARLO, ORLANDO
Era costui$_{P}$ quel paladin$_{N}$ gagliardo, figliuol d'Amon$_{PNNP}$, signor di Montalbano$_{NP}$, a cui$_{P}$ pur dianzi il suo destrier Baiardo per strano caso uscito era di mano. Come alla donna$_{N}$ egli$_{P}$ drizzò lo sguardo, riconobbe, quantunque di lontano, l'angelico sembiante e quel bel volto ch'all'amorose reti il$_{P}$ tenea involto.	This valiant knight$_{N}$ was Lord of Clarimount$_{NP}$, Duke Ammons$_{NP}$ sonne$_{NP}$, as you shall understand, Who$_{P}$ having lost his horse of good account, That by mishap was slipt out of his hand, He$_{P}$ followd him, in hope againe to mount, Untill this Ladies$_{N}$ sight did make him$_{P}$ stand, Whose$_{P}$ face and shape proportiond were so well, They seeme the house where love itselfe did dwell.	AMONE, ANGELICA, RINALDO
Mentre costei$_{P}$ conforta il Saracino$_{N}$, ecco col corno e con la tasca al fianco, galoppando venir sopra un ronzino un messaggier che parea afflitto e stanco; che come a Sacripante$_{PN}$ fu vicino, gli$_{P}$ domandò se con un scudo bianco e con un bianco pennoncello in testa vide un guerrier$_{N}$ passar per la foresta.	Thus while she$_{P}$ gives him$_{P}$ comfort all she$_{P}$ may, Behold there came a messenger in post, Blowing his horne, and riding downe the way, Where he$_{P}$ before his horse, and honor lost. And comming nearer he of them$_{P}$ doth pray, To tell if they$_{P}$ had seene passe by that cost, A champion$_{N}$ armd at all points like a knight, The shield, the horse, and armour all of white.	ANGELICA, BRADAMANTE, SACRIPANTE

identify all textual mentions referring to characters via nominal or pronominal forms (excluding extra-diegetic figures, collective groups, and non-human agents) and it must resolve every extracted instance to a specific character, returning the corresponding proper name alongside the exact textual mention. If a character's identity cannot be resolved, the model assigns the placeholder label `unknown_character`.

Narrative memory is treated as a binary variable (presence vs. absence: see § 4) to measure the impact of summary injection in mitigating the contextual fragmentation caused by breaking the poem into smaller textual units. This segmentation, in particular, introduces two risks of resolution failure: (a) *anaphoric assignment errors*, where models fail to link non-onomastic mentions to identities established in preceding batches; and (b) *cataphoric assignment errors*, where models cannot retrospectively resolve anonymous mentions whose identities are disclosed only in subsequent batches. These issues are especially pronounced in *Orlando Furioso*, where characters are referenced by diverse nominal expressions and the narrative structure is governed by *entrelacement* [19] – a technique that weaves together parallel plotlines, resulting in characters disappearing for long intervals and/or being introduced incognito, with their identification sometimes deferred until much later in the text.

4. Experiments

For our experiments, we established a **NER baseline** to serve as a reference point for LLM performance, utilizing two models: the CNN-based NER component of SpaCy's `it_core_news_lg` pipeline [20] and the Transformer-based `bert-italian-cased-ner`, fine-tuned for Italian NER [21, 22]; inference was conducted at the stanza level, retaining only entities classified as `PER` (Person). For **narrative memory generation**, different summaries were produced using two LLMs: `Llama-3.1-70B-Instruct` [23] and `Qwen-2.5-72B-Instruct` [24], with a fixed batch size of $n = 4$ stanzas[5]. For **character extraction**, we evaluated four LLMs: `Llama-3.1-8B-Instruct` [23], `Mixtral-8x7B-Instruct` [25], `Llama-3.1-70B-Instruct`, and `Qwen-2.5-72B-Instruct` (the latter three were loaded with 4-bit quantization)[6]; we tested multiple configurations by varying the batch size $n \in \{4, 8, 12\}$ and comparing runs with and without the inclusion of narrative memory[7]. **Evaluation** relied on standard quantitative metrics (Precision, Recall, F1 score) and qualitative error analysis (§ 5). To consider 16th-century orthographic variations and apocopated forms, extracted names were standardized (collapsing intra-stanza duplicates) and matched to ground-truth entries using the Ratcliff/Obershelp algorithm [26], which computes similarity based on shared substrings between two strings, with a threshold of 0.83[8]. We excluded `unknown_character` labels from quantitative metrics (as these often correspond to minor nameless characters absent from the ground truth), though they were retained for qualitative analysis.

5. Results and Error Analysis

As shown in Table 2, all LLM configurations (with a single exception[9]) outperform the NER baselines in terms of F1 score, largely driven by superior Recall. Furthermore, across all models, incorporating narrative memory yields a mean F1 increase of 8.1 percentage points. Regarding batch size, performance peaks at $n = 8$ but declines at $n = 12$, as excessive prompt length outweighs the benefits of additional context. The optimal configuration is `Qwen-2.5-72B` with $n = 8$ and `Qwen`-generated summaries.

To provide a more granular evaluation, we assess the precise impact of **narrative memory** on character identification (Table 3). Without narrative memory, the top-performing model (`Qwen-2.5-72B`) at $n = 8$ correctly identifies characters when their proper name is explicitly mentioned in the stanza (92.7% success rate) or when non-onomastic mentions follow it within the batch (79.3%), but performs poorly when mentions precede the first onomastic occurrence in the batch (23.0%) and inevitably fails when the proper name never appears within the batch (0.5%). Injecting `Qwen`'s summaries not only improves *intra-batch* resolution (rising to 86.6% for post-onomastic and 69.9% for pre-onomastic mentions) but also enables *inter-batch* resolution, albeit at a lower rate (achieving a 27.3% success rate for characters not explicitly named in the batch). The efficacy of narrative memory in resolving locally unnamed entities, however, is highly dependent on summary quality: in Canto IV, while `Llama`'s summaries fail to propagate the identity of `Bradamante` (unnamed until stanza 40), `Qwen` correctly maintains it via the previous summary from the end of Canto III and allows for a 32.6 percentage point improvement (from 19.6% to 52.2%). Furthermore, even if effective for *anaphoric assignment*, summaries cannot support *cataphoric assignment*, which explains a consistent performance drop in Canto IX (where F1 decreases to 0.59 in the optimal configuration), caused by the delayed onomastic denotation of characters in the Olimpia episode.

We now turn to a qualitative **error analysis** of the best-performing configuration, based on manual evaluation of the ten entities with the highest False Negative counts and the ten with the highest False Positive counts (Figure 2). The results are summarized in Table 4.

[5] Final canto batches may contain $m \leq n$ stanzas to consider cantos not divisible by n.

[6] We set the temperature to 0.1 and `max_new_tokens` to 1024.

[7] Models with unpromising results at $n = 4$ were excluded from trials at higher batch sizes, and `Qwen`-generated summaries were evaluated only on the top-performing configurations identified using summaries by `Llama`.

[8] In cases of multiple matches, the highest-scoring candidate was retained. The similarity computation was implemented using Python's difflib.SequenceMatcher.

[9] `Mixtral-8x7B` falls below the `bert-italian-cased-ner` baseline due to JSON formatting errors that affected the evaluation.

Table 2
Results across NER baselines and LLMs on Cantos I-XII. Configurations vary by batch size (*n*) and narrative memory (`Mem.`), with summaries (`Summ.`) from two LLMs. The highest F1 scores per group are in **bold**; the overall top performer is highlighted in gray. TNs represent *stanzas* where the model correctly identified no character mentions.
TP: True Positives; **FP**: False Positives; **FN**: False Negatives; **TN**: True Negatives; **P**: Precision; **R**: Recall; **F1**: F1 score.

Run Configuration				Counts				Metrics		
Model	**n**	**Mem.**	**Summ.**	**TP**	**FP**	**FN**	**TN**	**P**	**R**	**F1**
it_core_news_lg	-	-	-	482	691	1406	44	0.4109	0.2553	0.3149
bert-italian-ner	-	-	-	675	281	1213	85	0.7061	0.3575	**0.4747**
Mixtral-8x7B	4	No	-	468	323	1420	77	0.5917	0.2479	0.3494
Llama-3.1-8B	4	No	-	905	628	983	9	0.5903	0.4793	0.5291
Llama-3.1-70B	4	No	-	892	217	996	55	0.8043	0.4725	0.5953
Qwen-2.5-72B	4	No	-	1005	350	883	49	0.7417	0.5323	**0.6198**
Mixtral-8x7B	4	Yes	Llama-3.1-70B	719	407	1169	64	0.6385	0.3808	0.4771
Llama-3.1-8B	4	Yes	Llama-3.1-70B	1086	666	802	11	0.6199	0.5752	0.5967
Llama-3.1-70B	4	Yes	Llama-3.1-70B	1187	359	701	43	0.7678	0.6287	0.6913
Qwen-2.5-72B	4	Yes	Llama-3.1-70B	1308	420	580	55	0.7569	0.6928	**0.7235**
Llama-3.1-70B	8	No	-	1004	252	884	55	0.7994	0.5318	0.6387
Qwen-2.5-72B	8	No	-	1093	342	795	56	0.7617	0.5789	**0.6578**
Llama-3.1-70B	8	Yes	Llama-3.1-70B	1192	331	696	46	0.7827	0.6314	0.6989
Qwen-2.5-72B	8	Yes	Llama-3.1-70B	1354	344	534	58	0.7974	0.7172	**0.7552**
Llama-3.1-70B	8	Yes	Qwen-2.5-72B	1226	309	662	48	0.7987	0.6494	0.7163
Qwen-2.5-72B	8	Yes	Qwen-2.5-72B	1366	350	522	54	0.7960	0.7235	**0.7580**
Llama-3.1-70B	12	No	-	994	269	894	47	0.7870	0.5265	0.6309
Qwen-2.5-72B	12	No	-	1045	299	843	59	0.7775	0.5535	**0.6467**
Llama-3.1-70B	12	Yes	Llama-3.1-70B	1060	369	828	41	0.7418	0.5614	0.6391
Qwen-2.5-72B	12	Yes	Llama-3.1-70B	1238	326	650	56	0.7916	0.6557	**0.7173**

Table 3
TP and FN counts and Recall for `Qwen-2.5-72B` at $n = 8$: (left) across Cantos I–XII, without narrative memory vs. with narrative memory (`Qwen`-generated summaries); (right) in Canto IV, with `Llama`- vs. `Qwen`-generated summaries. **Mention Status**: `Onomastic` = first proper-name mention in the batch; `Pre-` / `Post-onomastic` = character mentions before/after the first onomastic occurrence; `Not in batch` = character mentions whose proper name never appears in the batch.

	Without Memory			With Memory		
Mention Status	TP	FN	Recall (%)	TP	FN	Recall (%)
Not in batch	2	431	0.5%	118	315	27.3%
Pre-onomastic	48	161	23.0%	146	63	69.9%
Onomastic	379	30	92.7%	377	32	92.2%
Post-onomastic	664	173	79.3%	725	112	86.6%

	Llama Summaries			Qwen Summaries		
Mention Status	TP	FN	Recall (%)	TP	FN	Recall (%)
Not in batch	9	37	19.6%	24	22	52.2%
Pre-onomastic	5	15	25.0%	14	6	70.0%
Onomastic	24	0	100.0%	23	1	95.8%
Post-onomastic	32	5	86.5%	33	4	89.2%

Regarding the top ten **False Negatives**, seven entities exhibit miss rates exceeding 50%: these errors stem from a *cataphoric assignment* failure and involve characters introduced anonymously, often within metadiegetic narratives[10] (`Atlante` and `Cimosco`) or as intradiegetic narrators[11] whose names are revealed only later (`Olimpia`, `Dalinda` and `Pinabello`); on average, these characters' proper names appear only after 40% of their total mentions, preventing correct resolution in earlier stanzas and jeopardizing performance even after name disclosure. The remaining three top-ten FNs correspond to major characters in the poem: collectively, they are affected in 49% of cases by undetected nominal references (such as "la bella donna", an expression shared by `Angelica` and `Bradamante`, and "il conte" or "il paladino" for `Orlando`) and in 39% of cases by undetected pronominal references, including enclitic forms (e.g., "promettendo*la*" at I.9.1, "tor*miti*" at VIII.74.8). Conceptually, their most critical omissions are the seven instances where a monoreferential noun phrase, which uniquely identifies the character, is overlooked by the model (e.g., "donna di Dordona" for `Bradamante` at XII.20.3, "cavallier di Brava" and "principe d'Anglante" for `Orlando` at VI.34.6 and XII.5.5, respectively).

[10]In Genette's terms, a metadiegetic narrative is a secondary narrative embedded within the primary diegesis [17].
[11]An intradiegetic narrator is a character situated within the diegesis who narrates at a subordinate narrative level [17].

Regarding the ten most frequent **False Positives**, hallucinations arise for two entities from improperly resolving minor nameless characters to common nouns (`re` and `eremita`) and in one entity from violating exclusion constraints (the steed `Baiardo`). More informative are the three cases linking FPs to the aforementioned FNs, where correctly detected occurrences are misresolved either as generic nouns (`Melissa` accounts for 12 of 16 instances of `maga`, and `Dalinda` for 9 of 13 instances of `donzella`, with the other 4 referring to `Olimpia`) or as honorific titles (`Cimosco` is successfully extracted 13 times but incorrectly labeled as `re di Frisa`)[12]. Finally, analysis of the remaining four top-ten FPs, all associated with character names, reveals two significant error types: 30% of these hallucinations consist of misattributions to other character entities (e.g., `Rinaldo` instead of `Polinesso` at V.12–16; `Orlando` instead of `Brandimarte` at VIII.89; `Alcina` instead of `Melissa` at VII.65 or `Logistilla` at X.43; `Ruggiero` instead of `Orlando` at XII.66–67) and 27% stem from implicit resolution, where the model infers a character's presence via verbal agency (e.g., `Rinaldo` at IV.72; `Orlando` at IX.6, IX.8, IX.91, XII.9, XII.79; `Alcina` and `Ruggiero` at VII.31–32; `Ruggiero` at X.71, X.92, X.107, X.113, XI.9) or via extended physical description (`Alcina` at VII.11–15) despite the absence of the required linguistic markers.

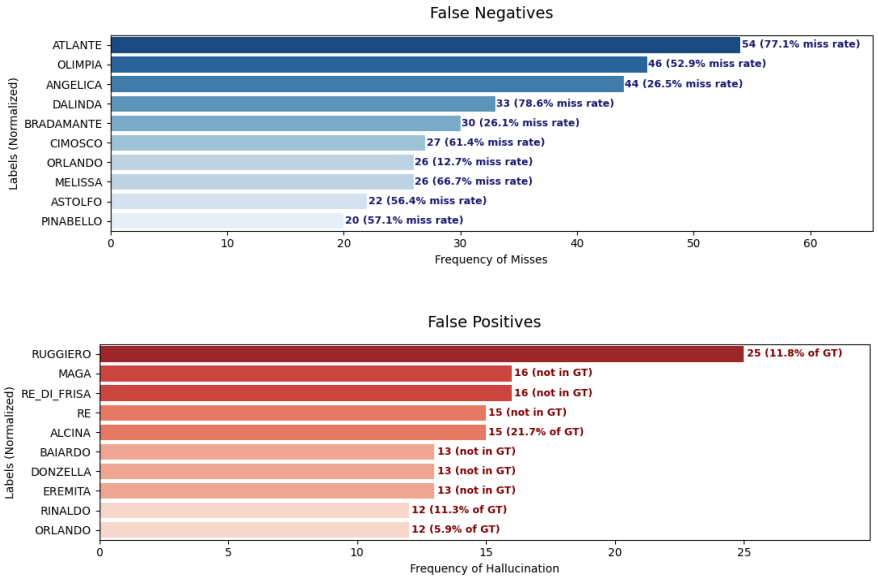

Figure 2: Errors for the optimal configuration, sorted by raw count. Upper panel: Top ten FNs, with miss rates calculated relative to each entity. Lower panel: Top ten FPs, showing hallucination rates relative to ground-truth occurrences (or marked "not in GT" for non-existent entities).

Table 4
Taxonomy of error types identified for the top-ten FNs and FPs in the optimal configuration. The **Ratio** reports the incidence of each error pattern against the total error count for the associated entities.

	Error Type	Ratio	Entities Evaluated
FN	Cataphoric Assignment Failure	129 / 208	`Atlante`, `Olimpia`, `Dalinda`, `Cimosco`, `Melissa`, `Astolfo`, `Pinabello`
	Undetected Nominal References	49 / 100	`Angelica`, `Bradamante`, `Orlando`
	Undetected Pronominal References	39 / 100	`Angelica`, `Bradamante`, `Orlando`
FP	Misresolution of Extracted Mentions	38 / 45	`maga`, `re di Frisa`, `donzella`
	Inclusion of Nameless Characters	28 / 28	`re`, `eremita`
	Character Misattribution	19 / 64	`Ruggiero`, `Alcina`, `Rinaldo`, `Orlando`
	Implicit Resolution	17 / 64	`Ruggiero`, `Alcina`, `Rinaldo`, `Orlando`
	Violation of Exclusion Constraints	13 / 13	`Baiardo`

[12]In related cases (though affecting only FN counts), correctly retrieved mentions are labeled `unknown_character` rather than being resolved to the appropriate entity (e.g., 18 instances for `Atlante` in Cantos II and IV).

Interestingly, these last patterns are echoed by an analysis of the **True Positives** in the optimal configuration, as over 10% (173/1366) of the extracted mentions associated with correct character identifications feature hallucinations: specifically, the model is prone to substituting atonic pronouns with tonic forms (e.g., "tu" for "ti"/"te" at IV.61, "ella" for enclitic "-la" at VIII.48.6) and to replacing pronominal references with the character's proper name (e.g., "Ginevra" instead of "ella" at V.9.2, "Bireno" instead of "lui" / "-lo" at IX.50) or with a plausible noun (e.g., "pagan" instead of "lo"/"gli" at I.66, "paladin" instead of "tu" at V.5.1). Such observations strengthen the suggestion that the model may draw on cues beyond explicit referring expressions when identifying character entities.

6. Limitations

The present study is preliminary. The dataset covers one fourth of *Orlando Furioso*, and the released version currently includes only resolved character names for each stanza. Since the annotation was performed by a single domain expert, inter-annotator agreement could not be calculated. However, the adopted annotation guidelines, based on objective linguistic markers, reduce interpretive ambiguity to a minimum. From a narratological perspective, a caveat (which nonetheless does not invalidate character identity extraction) is that explicit character mentions do not always coincide with character presence in the scene. Our evaluation was limited to open-source models; therefore, potential performance gains of proprietary models remain an area for future investigation. Finally, the NER baselines used here should be strictly interpreted as lower-bound reference points, as they are not fine-tuned for archaic poetic language and do not model long-range literary coreference.

7. Conclusion and Future Work

In this study, we presented a preliminary evaluation of LLM-based character identification applied to Italian Renaissance epics. Considering the complexities inherent in 16th-century poetic Italian, our experiments on *Orlando Furioso* yield encouraging results (F1 score of 0.758). The injection of narrative memory, in the form of LLM-generated summaries, proved effective in mitigating the impact of canto segmentation. Error analysis on the top ten FNs and FPs for the optimal configuration, however, shows that performance is still hindered by several factors, most notably the non-linear disclosure of character identities, which remains the primary driver of both missed detections and hallucinated resolutions.

We take these findings as a basis for future research. We plan to redefine the notion of character presence (currently tied to explicit nominal and pronominal markers) in terms of narrative agency — an operational shift that addresses LLMs' evident limitations in detecting pronominal references (often altered even in returned textual excerpts for TPs) and leverages the models' observed ability to implicitly infer character presence from their actions. Furthermore, we will prioritize the adoption (alongside or in place of summaries) of a structured narrative schema. This could take the form of a registry that tracks characters as anonymous entities (to be later identified through clustering and name resolution), associating them with stable properties (such as titles, lineage, and lordship, all recurrent traits in epic poetry), textual occurrences, and in-scene interactions with other entities.

As our work progresses, we may need to rely on ontological representations of narrative elements (most notably, at present, the GOLEM ontology [27]) to account for plot dynamics and for the distinction between the order of events in the story and their order in the narration (a phenomenon particularly relevant in cases of complex *entrelacement*). Our final aim is to advance the computational modeling of epic narratives and contribute to the development of digital literary scholarship.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.2 and Gemini 3 Pro in order to: Paraphrase and reword, Grammar and spelling check, Formatting assistance. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. Ranade, S. Dey, A. Joshi, T. Finin, Computational understanding of narratives: A survey, IEEE Access 10 (2022) 101575–101594. doi:10.1109/ACCESS.2022.3205314.

[2] B. Santana, R. Campos, E. Amorim, A. Jorge, P. a. Silvano, S. Nunes, A survey on narrative extraction from textual data, Artif. Intell. Rev. 56 (2023) 8393–8435. URL: https://doi.org/10.1007/s10462-022-10338-7. doi:10.1007/s10462-022-10338-7.

[3] V. Labatut, X. Bost, Extraction and analysis of fictional character networks: A survey, ACM Comput. Surv. 52 (2019). URL: https://doi.org/10.1145/3344548. doi:10.1145/3344548.

[4] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Transactions on Knowledge and Data Engineering 34 (2022) 50–70. doi:10.1109/TKDE.2020.2981314.

[5] R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak, The 8th international workshop on narrative extraction from texts: Text2story 2025, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 345–351.

[6] E. Gultepe, V. Mathangi, A quantitative social network analysis of the character relationships in the mahabharata, Heritage 6 (2023) 7009–7030. URL: https://www.mdpi.com/2571-9408/6/11/366. doi:10.3390/heritage6110366.

[7] C. Santini, L. Melosi, E. Frontoni, Named entity recognition in historical italian: The case of giacomo leopardi's zibaldone, in: Proceedings of the X-TAIL 2024 Workshop at the 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), volume 3967 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: https://ceur-ws.org/Vol-3967/X-TAIL-2024_paper_1.pdf.

[8] A. Belousova, J. S. Páramo Rueda, Macroanalysis of the strophic syntax and the history of the Italian Ottava Rima, Quantitative Approaches to Versification 72 (2019) 23–30. URL: https://versologie.cz/conference2019/proceedings/belousova-paramo-rueda.pdf.

[9] R. Delmonte, N. Busetto, Stress test for bert and deep models: Predicting words from italian poetry, International Journal on Natural Language Computing 11 (2022) 15–26. URL: https://aircconline.com/ijnlc/V11N6/11622ijnlc02.pdf. doi:10.5121/ijnlc.2022.11602.

[10] R. Sukthanker, S. Poria, E. Cambria, R. Thirunavukarasu, Anaphora and coreference resolution: A review, Information Fusion 59 (2020) 139–162. URL: https://www.sciencedirect.com/science/article/pii/S1566253519303677. doi:https://doi.org/10.1016/j.inffus.2020.01.010.

[11] Y. Liu, X. Peng, J. Cao, S, Y. Shen, T. Du, S. Cheng, X. Wang, J. Yin, X. Zhang, Bridging context gaps: Leveraging coreference resolution for long contextual understanding, in: Y. Yue, A. Garg, N. Peng, F. Sha, R. Yu (Eds.), International Conference on Learning Representations, volume 2025, 2025, pp. 64318–64334. URL: https://proceedings.iclr.cc/paper_files/paper/2025/file/a1b71d48d4806ecbe5a9e19fa3f10fdc-Paper-Conference.pdf.

[12] G. Martinelli, B. Gatti, R. Navigli, xCoRe: Cross-context coreference resolution, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 34264–34278. URL: https://aclanthology.org/2025.emnlp-main.1737/. doi:10.18653/v1/2025.emnlp-main.1737.

[13] D. Bamman, O. Lewke, A. Mansoor, An annotated dataset of coreference in English literature, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 44–54. URL: https://aclanthology.org/2020.lrec-1.6/.

[14] G. Martinelli, T. Bonomo, P.-L. Huguet Cabot, R. Navigli, BOOKCOREF: Coreference resolution at book scale, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vienna, Austria, 2025, pp. 24526–24544. URL: https:

//aclanthology.org/2025.acl-long.1197/.

[15] G. Genoni, Un grafo per il “furioso”. prime annotazioni di una ricerca sperimentale, AOQU (Achilles Orlando Quixote Ulysses). Rivista di epica 3 (2022) 241–290. URL: https://riviste.unimi.it/index.php/aoqu/article/view/18443. doi:10.54103/2724-3346/18443.

[16] L. Ariosto, Orlando Furioso, Biblioteca Italiana, 2003. URL: http://www.bibliotecaitaliana.it/scheda/bibit001301, digital version based on the print edition (Einaudi, Torino, 1966).

[17] G. Genette, Narrative Discourse Revisited, Cornell University Press, Ithaca, 1988. Originally published as *Nouveau discours du récit* (1983).

[18] S. J. Harington, Orlando Furioso in English Heroical Verse, Richard Field, for Iohn Norton and Simon VVaterson, London, 1607. Early English Books Online, University of Michigan Library Digital Collections. https://name.umdl.umich.edu/A21106.0001.001.

[19] F. Tomasi, Entrelacement, in: A. Izzo (Ed.), Lessico critico dell’Orlando furioso, Carocci, Roma, 2016, pp. 61–80.

[20] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spacy: Industrial-strength natural language processing in python, Zenodo (2020). URL: https://spacy.io. doi:10.5281/zenodo.1212303.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[22] Hugging Face Community, bert-italian-cased-ner, 2021. URL: https://huggingface.co/osiria/bert-italian-cased-ner.

[23] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[24] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, Z. Qiu, Qwen2.5 technical report, 2025. URL: https://arxiv.org/abs/2412.15115. arXiv:2412.15115.

[25] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mixtral of experts, 2024. URL: https://arxiv.org/abs/2401.04088. arXiv:2401.04088.

[26] J. W. Ratcliff, D. E. Metzener, Pattern matching: The gestalt approach, Dr. Dobb’s Journal 13 (1988) 46.

[27] F. Pianzola, L. Cheng, F. Pannach, X. Yang, L. Scotti, The golem ontology for narrative and fiction, Humanities 14 (2025). URL: https://www.mdpi.com/2076-0787/14/10/193. doi:10.3390/h14100193.