INDAQA2 - A Large Italian Narrative QA Benchmark: A CALAMITA 2026 Challenge

Luca Gioffré[1,*], Luca Moroni[1], Alberte Fernández-Castro[1,2], Elena Marafatto[1], Giacomo Garufi[1] and Roberto Navigli[1,2]

[1]*Sapienza NLP Group, Dip. di Ingegneria Informatica, Automatica e Gestionale, Sapienza University of Rome, Italy*
[2]*Babelscape, Rome, Italy*

Abstract

Long-context comprehension and reasoning remain largely underexplored in the evaluation of Italian Large Language Models (LLMs). Existing Italian benchmarks primarily focus on short or medium-length inputs, offering limited insight into models' ability to process extended narratives. To address this gap, we introduce INDAQA2, a substantially revised and expanded version of INDAQA, a benchmark for narrative question answering on original Italian literary texts. The new version comprises an expanded corpus of 461 total books, introduces a multiple-choice question answering format alongside the original open-ended tasks, and features manually curated texts drawn exclusively from works originally written in Italian, thus avoiding artifacts introduced by translation. The benchmark evaluates long-context understanding over complete books of up to 250K tokens, testing complementary comprehension skills through a dual-structure design: global narrative understanding, assessed via questions derived from book summaries, and local precision, assessed via questions grounded in specific passages and entity-level details. By supporting both open-ended and multiple-choice question answering formats, INDAQA2 enables evaluation of both generative capabilities and discriminative reasoning, facilitating comprehensive and scalable comparison across models. Our evaluation of several Italian-specialized and multilingual models reveals significant performance disparities across task formats and highlights limitations in how current Italian models utilize extended contexts.

Keywords

Narratives, Question-Answering, Long-context, Evaluation, Benchmark

1. Challenge: Introduction and Motivation

In recent years, an increasing number of newly released Large Language Models (LLMs) have been explicitly designed to process long inputs [1, 2]. Context windows have expanded rapidly, from early models limited to a few thousand tokens [3] to systems such as Llama 3.1 [4], supporting up to 128K tokens, and Qwen 2.5 [5] and Gemini 1.5 [6], which extend this capacity to the order of one million tokens. While these advances are impressive, the ability to handle long contexts introduces new and non-trivial challenges for model evaluation.

In parallel with the development of techniques to extend context length, there has been a surge in benchmarks targeting long-context capabilities. Prominent examples include Needle-in-a-Haystack–style tests [7, 8] and their long-context extension BABILong [9], which primarily assess retrieval and robustness under increasing input length. Beyond simple retrieval, benchmarks such as ∞Bench [10] extend evaluation to contexts exceeding 100K tokens across diverse domains, while RULER [11] introduces multi-hop tracing and aggregation tasks to test deeper reasoning capabilities. However, most of these benchmarks are strongly English-centric, with only a limited number of multilingual efforts (such as BABILong-ITA [12]), and they rarely focus on deep, narrative-level reasoning.

EVALITA 2026: 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Feb 26 – 27, Bari, IT
*Corresponding author.
gioffre@diag.uniroma1.it (L. Gioffré); moroni@diag.uniroma1.it (L. Moroni); castro@diag.uniroma1.it (A. Fernández-Castro); marafatto@diag.uniroma1.it (E. Marafatto); garufi.1750327@studenti.uniroma1.it (G. Garufi); navigli@diag.uniroma1.it (R. Navigli)
https://lukfre.github.io/ (L. Gioffré)
0009-0007-9705-6797 (L. Gioffré); 0009-0006-1210-5098 (L. Moroni); 0000-0003-3831-9706 (R. Navigli)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Despite recent efforts to advance generative evaluation for Italian [13, 14, 15, 16], a language spoken by over 60 million native speakers and supported by a rich literary tradition, comprehensive benchmarks for evaluating both retrieval and reasoning in long-document abstractive question answering remain largely absent [17]. The absence of such resources hinders the development and evaluation of Italian language models on tasks requiring extended discourse comprehension, a capability increasingly important for applications ranging from document analysis to educational tools and accessibility technologies.

To the best of our knowledge, there exist only two datasets for narrative QA in Italian, but each has significant limitations. FairytaleQA-IT [18], a machine translation into Italian of the FairytaleQA benchmark [19], is limited to children's fairy tales and may contain translation artifacts that do not reflect natural Italian language use. INDAQA [20], the first Italian long-context QA benchmark on narratives, adopts the NarrativeQA methodology of generating questions from human-written summaries rather than directly from source texts. While this approach ensures question quality, it may inadvertently constrain question types to those answerable from compressed versions of narratives, rather than questions requiring deep engagement with the full text. Moreover, the benchmark lacks a systematic quality check of the source texts, and by supporting only open-ended generative tasks, it requires a carefully tailored evaluation setup, both important factors for building a fair (narrative) benchmark [21, 22]. Supporting the multiple-choice task would allow for a more straightforward evaluation, despite its own structural problems [23].

To bridge this gap, we introduce an expanded and substantially refined version of INDAQA, which we name **INDAQA2**. The new benchmark addresses the aforementioned limitations and extends the dataset with an additional collection of 99 long texts paired with (open-ended and multiple-choice) question–answer items generated under different strategies, complementing the challenges and evaluation principles outlined in the CALAMITA initiative [16]. As in the original benchmark, all source texts come from gold Italian novels or theatrical screenplays, manually collected and curated. Importantly, the newly added works consist primarily of lesser-known literary texts, reducing the likelihood that models have memorized substantial amounts of content from their training data and thus mitigating contamination effects. Moreover, all texts are authored in Italian and written directly in the language, avoiding artifacts introduced by translation. In some cases, the texts also include regional varieties and dialects (e.g., works written by *Goldoni* or *Pirandello*), further increasing linguistic diversity. Together, these properties make the benchmark a more reliable and realistic test bed for evaluating long-context comprehension and reasoning in Italian LLMs. For the aforementioned reasons, we expect models to perform rather poorly on this new challenge, mainly due to the length of the inputs.

2. Challenge: Description

This challenge is focused on Open-ended (OE) and Multiple-choice (MC) QA over long, narrative texts. Thus, it supports two main tasks, depending on the expected output of the model (free-form, open-ended generation or choice labels).

We focus on narratives for two main reasons. First, **many narrative works are long by-design**, so we can avoid constructing a synthetic benchmark stitching together documents coming from different sources, as done in RULER [11] and BABILong [9]. The second reason is that **the narrative domain represents a critical challenge** for natural language understanding systems [21]. Unlike factoid QA tasks over short passages [24, 25], the narrative QA task requires models to maintain coherence across thousands of tokens, track character relationships, understand causal chains, and reason about plot developments.

All QA items in the benchmark have been automatically generated either from the summary of the story, following the style of NarrativeQA [26], or from individual or grouped excerpts (Section 3).

INDAQA2 - OEQA task The model is prompted with a simple instruction, followed by the entire content of the book and the question. Then, the model is expected to generate a concise answer, which will be evaluated against a set of reference answers. The number of reference answers, which spans

from 1 to 5, depends on the difficulty of answering the question: questions which may allow for different formulations of the correct answers (i.e., paraphrases) have more references, so that the generated answers are scored fairly.

INDAQA2 - MCQA task The model is prompted with a simple instruction, followed by the entire content of the book, the question, and four answer choices (three distractors and the correct answer). The choices are already shuffled, so that the correct answer can be found in any position with equal probability. The model is expected to identify the correct answer and return the letter (i.e., the label) corresponding to the option deemed correct. Since the correct answer is given as one of the four options, in this setting there is no need for a set of reference answers.

```
la_coscienza_di_zeno.oe.42

Chi è Carla per Zeno?
- È la sua amante
- È la donna con cui ha una
 relazione extraconiugale.
```

```
la_coscienza_di_zeno.mc.42

Chi è Carla per Zeno?
A. È un'amica di sua moglie
B. È una parente
C. È la sua amante
D. È la sua cameriera
```

Figure 1: Examples of a QA item from *La coscienza di Zeno*. In the left side is the OE item, for which both reference answers are true. In the right side, is the MC item: only one of the four options is correct (marked in **bold**).

3. Data description

Our goal is to create a large-scale question-answering benchmark grounded in Italian literary texts. Building upon the 362 books from the original INDAQA, we extend the collection with 99 additional works, resulting in a total of 461 books sourced from Project Gutenberg, Wikisource, and LiberLiber (Section 3.1). We also collect introductory sections as well as summary or plot sections from the corresponding Wikipedia pages when available. However, manually writing questions, reference answers, and distractors (collectively, QA items) at book scale is unfeasible. Therefore, we rely on an LLM to generate and refine QA items, following an approach similar to [20].

Our QA generation is grounded in different types of textual context, which leads to two data splits. In the **Summary-level** split, QA items are generated from book summaries available on the corresponding Wikipedia pages (Section 3.2). In the **Passage-level** split, QA items are instead generated from individual book passages, as no summaries are available for these works (Section 3.3).

To assess the quality and appropriateness of the generated data, we conduct a manual annotation study on a representative sample of QA items (Section 3.4).

The data format and the inference prompts used in our evaluation are described in Sections 3.5 and 3.6. We provide detailed statistics in Section 3.7 and Section C, while the prompts used for QA generation and refinement are available in Section A.

3.1. Origin of data

All documents in the benchmark have been downloaded from either Project Gutenberg, Wikisource or LiberLiber[1]. We also use Wikipedia to extract summaries and metadata (when available, see Section B).

Project Gutenberg is a volunteer-driven initiative dedicated to digitizing and preserving cultural works and to encouraging the creation and distribution of e-books. Founded in 1971, it is the oldest existing digital library. Its collection consists mainly of complete books or individual texts in the public

[1] https://www.gutenberg.org/, https://it.wikisource.org/, https://liberliber.it/

domain. All materials are freely accessible and provided in open, non-proprietary formats compatible with nearly all computing platforms.

Wikisource is a project of the Wikimedia Foundation that aims to build a freely accessible online library of original source texts and their translations in any language. Unlike Project Gutenberg, Wikisource documents can be collaboratively edited and reviewed by contributors, allowing continuous improvement, verification against original sources, and transparent version tracking. As a result, the quality and reliability of texts generally increase over time.

LiberLiber is an Italian non-profit project focused on promoting free culture through the publication of literary, musical, and scholarly works in the public domain or released under free licences. Its digital library places great emphasis on textual accuracy, careful proofreading, and scholarly reliability. Often, documents hosted on LiberLiber are considered to be of higher editorial quality than those available on Project Gutenberg.

3.2. Summary-level data split

Source texts The Summary-level data split contains 362 books for which a summary or plot section is available on Wikipedia, corresponding to the original INDAQA dataset [20]. During our analysis, we found that several books contained incomplete or corrupted content. For instance, many theatrical screenplays were missing character names in dialogues, while others exhibited missing or duplicated sections. These issues were likely caused by failures of the HTML parser when applied to heterogeneous, crowdsourced texts that lack a standardized structure. It is indeed difficult to devise a parsing method which accounts for different HTML formatting, as one effective for one document could produce corrupted outputs for others. To address this issue, we manually downloaded and reviewed the source texts for all documents, ensuring high quality across the dataset[2].

QA items generation Since the QA items were generated from the summaries, which we found to be of high quality, and manually validated in [20], we do not modify them.

The resulting Summary-level section of INDAQA2 contains the original summaries, questions and answers, with clean source books for each QA item. Figure 13 in Section D shows an example of a Summary-level QA item as a `JSON` object.

3.3. Passage-level data split

Source texts While valuable, the Summary-level data split mainly accounts for questions either about the whole narrative (*abstractive*) or about information for which the answer can be found in various passages. To test the ability of the models to retrieve more localized and specific details, we extend the original 362 documents with a collection of another 99 books that were discarded in the previous INDAQA release since they were lacking a summary. These books were downloaded from the same sources and followed the same approach of the aforementioned data split.

QA item generation Lacking a summary, we devised three methods for generating QA items with an LLM starting from individual passages. These methods categorize the questions into three sets:

1. **Local**: questions generated from a single passage (defined as 20 contiguous sentences) randomly selected at runtime (prompt in Figure 5). These questions typically focus on specific details stated in the provided passage.
2. **Local (alternative)**: questions generated from a single passage plus the previously generated *Local* items (prompt in Figure 6). We noticed that this generation setting encourages the model to avoid repetition and generate less straightforward questions, shifting the distribution of question types (see Figure 3).

[2]We downloaded the text directly in `.txt` or `.rtf` formats, avoiding HTML pages.

3. **Entity**: questions generated from three passages in which an entity consistently appears, selected from the beginning, middle, and ending sections of the documents (prompt in Figure 7). Entities are identified by extracting all capitalized names (excluding *stopwords*) from the questions in the two previously defined sets. We clustered the entities by a simple exact string match. These capitalized terms capture recurring entities such as characters, locations, or organizations that are central to the narrative. For each entity cluster, we select three passages where the entity appears. Entities with fewer than three occurrences or very low frequency are filtered out. The resulting clusters are manually validated to remove noise (e.g., common words erroneously capitalized, non-entities). Finally, the selected passages and their associated local questions are provided to the LLM to generate questions targeting overarching plot elements, character development, or thematic connections across the narrative.

MCQA conversion Due to their localized scope (i.e., a single passage), *Local* and *Local (alternative)* QA items are well suited to a multiple-choice format. For each item in these two sets, we provide an LLM with the source passage and the generated question with reference answers, and instruct it to produce three plausible distractors, preferably grounded in the given context (prompt in Figure 8). We also tried to convert the *Summary* and *Entity* questions to a multiple-choice format. However, we found that generating plausible distractors from summaries or multiple distant passages resulted in hallucinations or generally weak, low-quality distractors. Hence, only the two *Local* sets support the MCQA format.

QA items refinement Following the original methodology for the INDAQA dataset, we refine the *Local* QA items by asking an LLM to assess whether the questions are well posed and answerable, and whether the corresponding answers are acceptable. We exclude *Entity* questions from this correction step because their higher complexity could not be reliably handled by the tested LLMs. Figure 9 shows the prompt we used for this task, which we built following guidelines similar to those of [21]. The refinement process identified 49 QA items with problems: given their low number, we manually review the LLM correction and substitute the refined QA items in the benchmark.

Throughout our experiments, we used Gemini-2.5-Flash [27] to generate, refine and convert QA items. An example QA item for each question set can be found in Section D, Figures 14 to 16.

3.4. Annotation details

To assess the quality and validity of the generated QA items, we conduct human validation on a representative sample of the corpus. We focus on the newly added documents[3], following similar annotation guidelines as in the original INDAQA [20].

Subset Selection The Passage-level split comprises 11,560 QA items across 99 books, of which 10,187 support both free-form and multiple-choice evaluation (the two *Local* question sets), while the remainder consists of only open-ended questions (the *Entity* question set). We target approximately 5% of the total dataset for human annotation. We adopt a stratified sampling strategy to select the books in the annotation set. Books are divided into 20 equal-probability bins based on text length quantiles, and from each bin, the book whose length is closest to that bin's mean length is selected (Figure 11). We focus on 20 bins (and so, 20 books) to ensure manageable scope while maintaining diversity. This approach ensures representative coverage across the entire range of text lengths, avoiding potential biases.

Then, we randomly sample the QA items from the set of 20 books to mirror the distribution of the three question sets: 400 items from *Local* set (20 items per book), 120 items from *Local alternative* set (6

[3]The original INDAQA has an error rate of the generated QA items of 2.32%, which we deem acceptable.

items per book), and 60 items from *Entity* set (3 items per book), yielding a total of 580 elements to review.

To measure inter-annotator agreement (IAA), we randomly select 100 overlapping items from the annotation subset (~17%). This overlap also mirrors the overall QA item distribution: approximately 70% from the *Local* set, 20% from the *Local (alternative)* set, and 10% from the *Entity* set.

Annotation Guidelines The items in the annotation sample were independently validated by two expert annotators (either native or proficient in Italian). Annotators were asked to assess the quality of automatically generated QA items for Italian long-context narratives. For each of them, annotators evaluated the following dimensions:

- **Fluency**: Whether the question, correct answers, and eventual distractors are grammatically correct and naturally phrased in Italian.
- **Validity**: Whether the item elements are appropriate and accurate:
  - The question is clear and answerable given the source text
  - The reference answers are factually correct
  - The distractors are plausible but incorrect (if present)

Annotators were encouraged to note any unusual patterns, ambiguities, or issues worthy of further investigation.

Result After the annotation phase, we observed high inter-annotator agreement, with a Cohen's Kappa of 0.7563. The average error rate in the dataset, defined as the proportion of non-acceptable items among the annotated sample, is 4.74%. From the annotation process, we observed that the generated questions were always well-posed and answerable from the provided context alone. The errors focused mostly on the reference answers, in particular on the second reference, which was sometimes inaccurate or wrong, while the first was always correct. Regarding the distractors, they were generally plausible, albeit sometimes weak. We discuss the implications of these quality observations for our evaluation methodology in Section 6.

3.5. Data format

The benchmark is freely available through the Hugging Face repository[4]. All items have the same data fields, but depending on the data section and question set, some may be empty.

- **id** (`str`) : unique identifier for the document
- **text** (`str`) : text of the document
- **qas** (`list[dict]`) : QA entries associated with the document
  - **question_id** (`str`) : unique ID for the QA item
  - **question** (`str`) : the question text
  - **answers** (`list`) : list of free-form reference answers
  - **choices** (`list`) : list of MCQA options
  - **target** (`dict`) :
    - **label** (`str`) : correct MCQ label (`'A'`, `'B'`, `'C'`, or `'D'`)
    - **text** (`str`) : canonical correct answer (i.e., the first reference)
  - **entity** (`str`) : entity targeted by the question, if present
  - **model** (`str`) : generator model used
  - **kind** (`str`) : question type (*Local*, *Local Alternative* or *Entity*)
  - **source_paragraphs_ids** (`list`) : list of paragraph indices used to generate the QA
  - **source_questions_ids** (`list`) : list of related question indices
- **metadata** (`dict`) : book-level metadata

[4] `https://huggingface.co/datasets/sapienzanlp/INDAQA_CALAMITA`

Metric	Summary-level set	Passage-level set	Both sets
# Documents	362	99	461
# QA items	13,661	11,560	25,221
# QA items/doc	38 ± 2	117 ± 20	55 ± 34
Text length average	26K ± 33K	58K ± 31K	33K ± 35K
Text length range	0.5K-242K	8K-188K	0.5K-242K

Table 1
Statistics for the documents in INDAQA2 divided by data split.

- **title** (`str`) : title of the work
- **author** (`str`) : author name
- **year** (`int`) : publication year
- **summary** (`str`) : book summary used in summary_level
- **summary_length** (`int`) : length of the summary (in words)
- **text_length** (`int`) : length of text (in words).
- **source_link** (`str`) : link to the text source
- **summary_link** (`str`) : link to the summary source
- **qa_paragraphs** (`list[str]`) : list of text chunks used to generate the QAs

We also report one example per data split and question set in Appendix Section D, Figures 13 to 16.

3.6. Example of prompts used for zero or/and few shots

The challenge is supposed to be accomplished in a zero-shot setting with a very simple prompt. We show in Figure 4 the prompt used for inference (in the generative task, `choices_block` is null).

3.7. Detailed data statistics

Tables 1 and 2 report statistics for the Summary- and Passage-level splits of the INDAQA2 dataset. Specifically, Table 1 presents document-level statistics divided by data split, including the number of documents, the average number of QA items per document, and text length. Table 2, instead, summarizes the distribution of question types (Summary, Local, and Entity) across the dataset. In Figure 2 we plot the length distribution of both data splits. Additionally, in Figure 3 we plot the question categorization (by first word) distribution across question types. Further statistics are presented in Section C.

Summary-level This split comprises 13,661 open-ended QA items spanning 362 documents, averaging approximately 38 QA items per book. Due to the cleaning process applied to the original INDAQA dataset, document lengths were reduced by about ~3% while preserving all narrative content, from an average of 27K ± 38K words to an average of 26K ± 33K words.

Passage-level This split comprises 11,560 open-ended QA items spanning 99 documents. While containing fewer documents, it still has a comparable number of QA items with respect to the Summary-level split through a higher density of QA items. Notably, Passage-level documents are more than double the length of *Summary-level* ones on average, requiring models to process and reason over more extensive textual contexts.

4. Metrics

We employ different evaluation methodologies for the two settings in INDAQA2: *n*-gram-based metric for generative questions and accuracy for multiple-choice questions. In practice, as with all CALAMITA challenges, evaluation is carried out using the LM Evaluation Harness framework developed by EleutherAI.[5]

[5] https://github.com/EleutherAI/lm-evaluation-harness

Question Type	# Items	# Items/doc	Question length	Answer length
Summary-level split				
Summary	13,661 (100%)	38 ± 2	7 ± 2	5 ± 3
Passage-level split				
Local	7,901 (68%)	80 ± 14	8 ± 2	4 ± 2
Local (alternative)	2,286 (20%)	23 ± 5	9 ± 3	6 ± 4
Entity	1,373 (12%)	14 ± 6	13 ± 3	24 ± 8

Table 2
QA item distribution and length statistics by question type. Entity Questions feature notably longer answers (average 24 words vs. 4–6 words for other types), while Local Questions dominate the Passage-level set (68%). Question and answer lengths are measured in words; percentages represent proportions within each set.

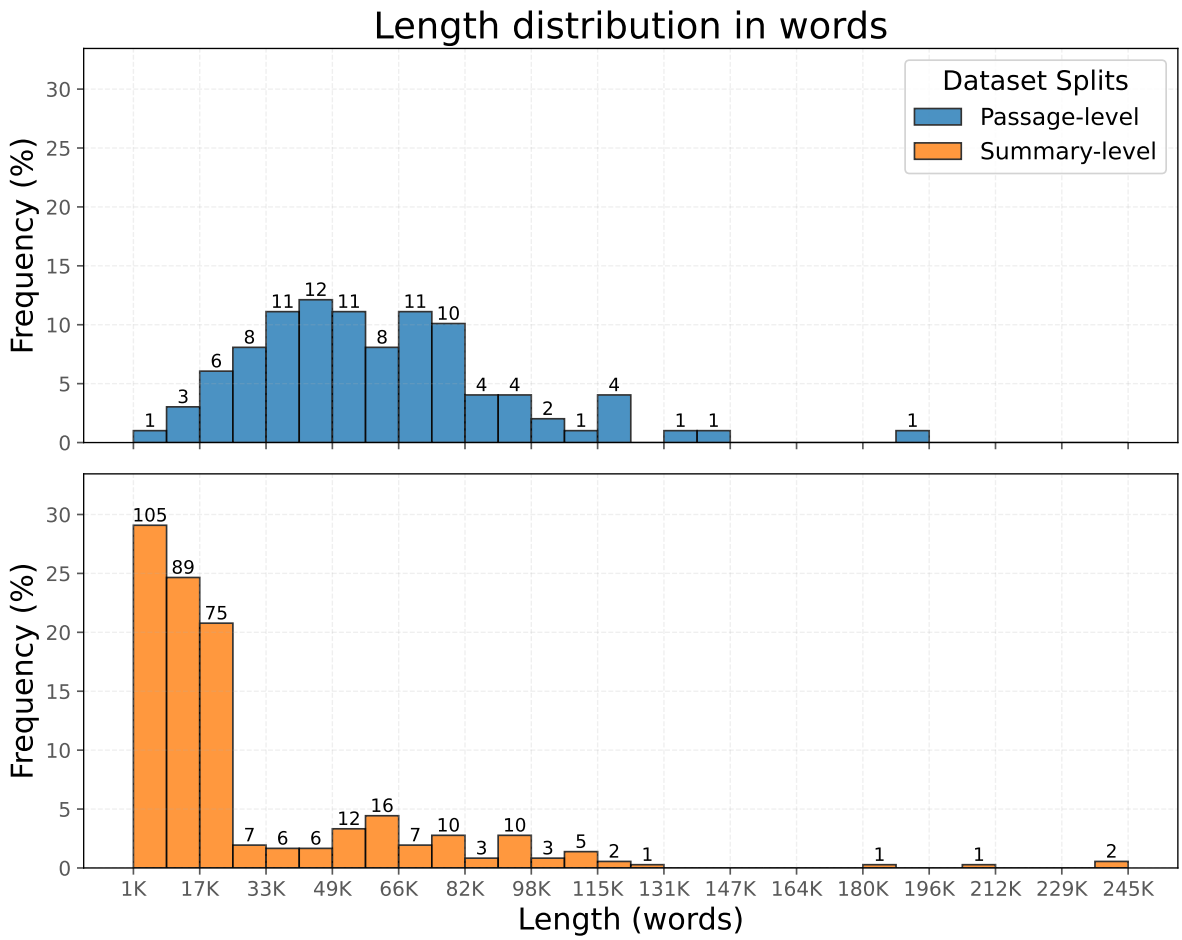

Figure 2: Document Length distribution across the two splits of INDAQA2. While the distribution of the original dataset is skewed towards shorter documents, the new data split is more balanced.

OEQA Evaluation For the generative task of INDAQA2, we employ **Exact-Match** (EM) and **METEOR** [28, 29] as automatic evaluation metrics. For both metrics, we take the maximum score between the candidate answer and the set of reference answers as the score of the model for that QA item.

EM is a simple and intuitive measure that checks whether a reference answer appears verbatim within the generated output. However, this metric may fail to capture semantically correct responses that do not exhibit lexical overlap with the reference. To address these well-known limitations of EM, we additionally adopt METEOR.

METEOR computes an alignment-based score between generated and reference answers by considering exact, stem, synonym, and paraphrase matches. Unlike simpler n-gram-based metrics such as ROUGE, METEOR jointly

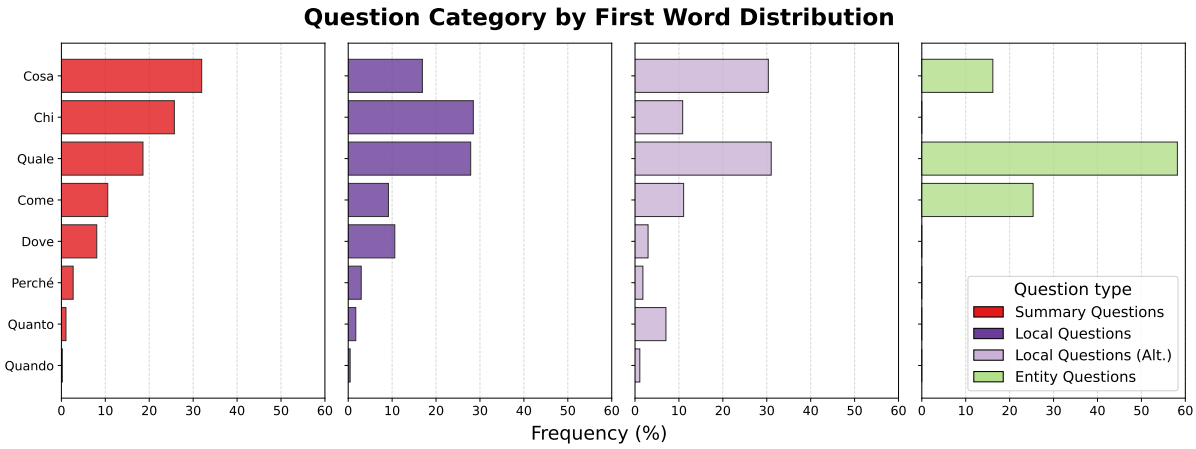

Figure 3: Distribution of first words of the questions. Each set has a different percentage of questions starting with a given first word.

models precision and recall while leveraging linguistic knowledge (e.g., synonyms) from WordNet, making it particularly suitable for evaluating natural language answers with different but equivalent surface forms. While there is not a standard metric for the evaluation of open-ended QA tasks, we follow prior work that demonstrates METEOR's superior reliability in this domain [21].

MCQA Evaluation For the MCQA component of INDAQA2, we follow best practices for multiple-choice evaluation proposed in recent works, in which model outputs are evaluated using manually curated regular expressions [23, 30]. The set of regex patterns is constructed by inspecting the outputs of a subset of models and identifying the most frequent surface forms used to indicate the selected option. This approach follows the analysis and setup of [23], which shows that regex-based answer extraction is more reliable than perplexity-based evaluation pipelines that rely on next-token probabilities over candidate choices. Accuracy is computed as the proportion of questions for which an extracted pattern corresponds to the correct answer. The complete list of regex patterns is provided in Appendix E.

Context Truncation INDAQA2 contains documents of highly variable length, reaching up to hundreds of thousands of words (see Figure 2), which make the benchmark both memory- and computation-intensive. To have an affordable evaluation, we define multiple context-size settings by truncating each book to its first words[6]. Specifically, besides evaluating on the full book, we also test the models under two context lengths: 10K and 50K words. This setup allows us to assess system performance not only under partial but manageable contexts, but also in terms of efficiency in word splitting. Models that require fewer splits (i.e., having a lower fertility) are better suited to incorporate partial or full contextual information within their context window.

5. Results, Challenge-Specific Insights and Lessons Learned

We analyze and report the results of the INDAQA2 challenge, examining the performance of LLMs featured on the CALAMITA leaderboard. Our analysis is complemented by a qualitative discussion of a sample of generated outputs, as well as a critical assessment of current limitations and future challenges in Italian narrative understanding.

Performance across models (types, sizes) Table 5 reports the performance of the five instruction-tuned models evaluated in the INDAQA2 challenge. Specifically, we consider two Llama models—Llama3.1-8B and Llama3.1-70B [4]—the ANITA-8B model [31], and two variants of Minerva: Minerva-7B [32] and Minerva-7B$_L$ [20]. Minerva-7B$_L$ is an extension of Minerva-7B that underwent continual training to increase the maximum context length from 4K to 32K tokens. By comparison, ANITA-8B natively supports an 8K token context window, while both Llama models handle input sequences of up to 128K tokens.

[6]We simply split on whitespace.

QA item	Model answers	METEOR
Question Dove stava andando la marchesa?	*ANITA-8B*: Non andava da nessuna parte, stava seduta in un salotto di Bologna.	0.31
References – A Bologna – Verso la città di Bologna	*Llama-8B*: La marchesa stava andando a pranzo da Giorgio, il conte De Vinci, e poi avrebbe accompagnato Mimy, la moglie di Carlo, a pranzo da lei.	0.13
	Minerva-7B$_L$: La marchesa stava andando a Bologna.	0.89

Table 3
Model responses on a Local question in the OE task, paired with their METEOR score.

Among the 7-8B models, performance patterns vary considerably across context sizes and task types. In the 10K-word context setting, Llama3.1-8B achieves the strongest results on the *Local* and *Local alternative* multiple-choice tasks, reaching 43.83% and 43.15% accuracy respectively. Interestingly, the pattern reverses for the more challenging open-ended generation tasks: Minerva-7B$_L$ outperforms all other 7-8B models on the *Summary*, *Local*, and *Entity* sets when measured by METEOR and EM scores. This suggests that Italian-specialized models may retain an advantage for generative tasks requiring nuanced language production, even when the multilingual model are better at discriminative multiple-choice selection. Notably, Minerva-7B$_L$ consistently surpasses both the English-centric Llama models and the other Italian-specific alternatives (ANITA-8B and the previous version Minerva-7B) on these open-ended tasks.

When context is extended to 50K words, Llama3.1-8B and Llama3.1-70B show large improvements on all the tasks, due to their capabilities in handling longer contexts. This is especially evident in the MC task, with an increment of more than 20 percentage points. In contrast, the corresponding METEOR scores for the open-ended *Local* set increase by only 12 points. This pattern suggests that Llama models are particularly adept at leveraging extended context for multiple-choice tasks, where the presence of answer options provides additional grounding that helps the model locate and select the correct response. For open-ended generation, however, the benefits of longer context are more modest: while the model may identify relevant information, accurately producing well-formed Italian answers remains challenging.

Italian-specialized models show minimal improvement on 50K words and full book settings across both task formats, with ANITA and Minerva variants demonstrating performance nearly identical to their 10K results. This stagnation is due to their limited context window with respect to Llama models.

With full book context available, Llama3.1-8B maintains its strong performance while showing only incremental gains over the 50K setting. This plateau suggests that 50K words may already encompass most information relevant to answering the benchmark questions, with additional context providing diminishing returns.

As expected, Llama3.1-70B substantially outperforms all smaller models across all metrics and context settings, with performance gains becoming more pronounced as context length increases. The model's higher parameter count enables it to better exploit the additional information provided by longer contexts. Due to computational constraints, we were unable to evaluate Llama3.1-70B on the full book setting.

Error Analysis To further investigate the task-specific performance patterns observed in the quantitative results, we conduct a qualitative error analysis on selected examples from both evaluation settings. Table 3 illustrates model outputs for question from the *Local* set in the OE task. The results demonstrate that, in this example, METEOR scores effectively capture answer quality: Minerva-7B$_L$ produces a more general but correct response, achieving a higher METEOR score. In contrast, ANITA and Llama-8B generate less accurate or incomplete answers, reflected in their lower scores. This example corroborates our earlier finding that Italian-specialized models maintain advantages in open-ended generation tasks when they can effectively access relevant context.

Table 4 presents a particularly revealing case where Minerva-7B$_L$ is prompted with the same question under both OE and MC formats. While the model successfully generates the correct answer in the OE setting, it selects an incorrect option in the MC setting. This failure mode is consistent with the quantitative patterns discussed above: Minerva-7B$_L$ demonstrates good performance on open-ended tasks but struggles with multiple-choice selection. The contrast highlights a fundamental difference in how these evaluation formats probe model capabilities, a finding that aligns with observations by [20]. While multiple-choice tasks primarily assess whether models can leverage answer options to discriminate between alternatives (a capability where Llama models excel), OE tasks require models to independently generate well-formed responses in the target language, where Italian-specialized models show relative strength.

QA item	Model answers	Score
Question Da quale città provenivano le commissioni per estirpare l'eresia valdese? **Choices** A. Da Carignano B. Da Ginevra C. Dalla Francia **D. Da Torino**	*OE setting* Da Torino *MC setting* B. Da Ginevra	METEOR: 1.00 Correct: False

Table 4
Comparison of the Minerva-7B$_L$ responses on a Local question in the OE task and MC task, paired with their evaluation. In the OE task, the model correctly responds, while in the MC task, it is confounded by the four options and chooses the wrong one. The choice in **bold** is the correct response, which corresponds to the one of the reference answers in the OE task.

Model	Summary set		Local set			Local alt. set			Entity set
	EM	METEOR	EM	METEOR	ACC	EM	METEOR	ACC	METEOR
10K-word context									
ANITA-8B	3.92	23.12	10.38	14.92	38.08	4.16	11.47	36.37	20.15
Minerva-7B	3.24	25.37	8.15	16.98	28.57	2.89	12.56	25.08	21.56
Minerva-7B$_L$	**5.08**	**26.93**	11.87	**20.31**	33.52	6.08	15.31	27.48	**22.21**
Llama3.1-8B	4.46	25.60	**14.44**	18.66	**43.83**	**9.89**	**16.65**	**43.15**	20.83
Llama3.1-70B	3.41	<u>27.62</u>	<u>17.04</u>	<u>23.24</u>	<u>49.57</u>	<u>11.72</u>	<u>19.71</u>	<u>50.02</u>	20.01
50K-word context									
ANITA-8B	4.26	24.06	11.37	15.36	39.46	4.86	11.79	37.64	20.57
Minerva-7B	3.45	25.81	8.37	17.02	29.91	2.97	12.64	25.95	21.46
Minerva-7B$_L$	5.30	27.80	13.10	21.05	33.91	5.99	15.07	26.43	22.17
Llama3.1-8B	**5.45**	**28.42**	**29.44**	**30.60**	**64.98**	**26.20**	**34.16**	**66.91**	**22.87**
Llama3.1-70B	3.21	<u>29.38</u>	<u>36.22</u>	<u>41.75</u>	<u>78.80</u>	<u>32.94</u>	<u>45.23</u>	<u>80.17</u>	<u>24.35</u>
Full book context									
ANITA-8B	4.21	25.38	10.99	15.51	39.50	4.72	12.04	37.68	19.55
Minerva-7B	3.61	26.28	7.68	16.71	30.42	2.62	12.34	26.12	21.31
Minerva-7B$_L$	5.42	27.94	12.94	20.75	33.54	5.90	15.11	26.47	22.10
Llama3.1-8B	**5.51**	**30.74**	**31.73**	**34.24**	**67.77**	**28.39**	**37.87**	**70.15**	**22.77**
Llama3.1-70B	-	-	-	-	-	-	-	-	-

Table 5
Results of four models on INDAQA2. We highlight in **bold** the best result per context setting on 7-8B models, and we <u>underline</u> the best result considering also the 70B model. Due to the high length of the responses in the Entity set, we do not show EM scores, as they are close to zero.

Critical analysis/discussion/future Our results carry important implications for the development of Italian language models and long-context evaluation more broadly. Current Italian-specialized models struggle with extended context utilization, suggesting that architectural innovations beyond continued pretraining may be necessary. Future work should explore whether techniques such as modified attention mechanisms, retrieval-augmented approaches, or different positional encoding schemes can better enable smaller Italian models to leverage long contexts. Moreover, the disparity between MC and OE performance raises questions about evaluation methodology: while both formats provide valuable signals, developing evaluation frameworks that better disentangle retrieval capabilities from generation quality could yield more actionable insights for model development.

6. Limitations

Historical and Linguistic Context In order to be copyright-free, INDAQA2 contains Italian literary texts spanning the period 1827-1948. As such, the language employed in these texts reflects the lexical and grammatical conventions of 19th and early 20th century Italian, including archaic vocabulary and syntactic structures that may differ substantially from contemporary usage. Additionally, the vast majority of authors in the corpus are males, which mirror the gender disparities inherent in that period. Consequently, the narrative content may embody attitudes, ideologies, and perspectives that reflect the sociocultural aspects of the period, some of which diverge from contemporary ethical and social standards.

Synthetic Data Although the QA items underwent quality control procedures, it is important to note that they were generated using an LLM. As such, the dataset may exhibit certain limitations inherent to LLM-generated content, including potential factual inaccuracies and systematic biases that reflect the model's training data and architectural characteristics. Our annotation process (Section 3.4) revealed that while questions were consistently well-posed and answerable, approximately 4.74% of items contained errors, primarily in secondary reference answers. Since our evaluation setup (Section 4) computes the maximum of EM or METEOR scores across all references, the presence of at least one correct reference ensures valid answers are properly credited. However, inaccurate secondary references could theoretically reward incorrect model responses that happen to align with those errors rather than the correct answer. Given the low error rate and the fact that such biases manifest *only* when a model's output matches the wrong reference instead of the correct one, we expect the practical impact on benchmark results to be limited. Nevertheless, future work could involve manual correction of identified errors to further enhance dataset quality.

Evaluation Setup While METEOR provides a reasonable automated evaluation framework for our task, we acknowledge its limitations. Automatic metrics based on surface form similarity may not fully capture semantic correctness, especially for questions requiring reasoning or inference. An LLM-as-a-Judge framework, where a powerful language model evaluates answer quality according to rubrics, has been shown to yield scores more aligned with human judgments in recent work [21, 33]. However, such approaches introduce additional computational costs, potential biases from the judge model, and complications in reproducibility. For the purposes of establishing baseline performance and enabling rapid iteration, we consider METEOR a pragmatic choice that balances evaluation quality with practical constraints.

7. Ethical issues

Given the publication years of the dataset (1827-1948), the source texts reflect the historical and sociocultural context of their time and may contain biased, stereotypical, or otherwise sensitive portrayals related to gender, ethnicity, religion, violence, or other forms of toxicity. While these aspects are intrinsic to the literary material and are preserved for research fidelity, their uncritical use may lead to the reproduction or amplification of outdated viewpoints in downstream applications.

We do not attempt to modify or filter such content, and we encourage users to interpret the dataset within its historical context and not to train models nor deploy models trained on it. The dataset is intended **exclusively** for research and benchmarking purposes.

8. Data licence and copyright issues

The dataset is constructed from publicly available data sources. The synthetic QA items do not have a copyright, as they respect the provider usage policy[7]. We will release the dataset upon acceptance for research use. All original copyrights remain with the respective content owners, and the dataset does not redistribute proprietary or restricted material. Users of the dataset are responsible for ensuring compliance with the licences of the original sources when using the data.

Acknowledgments

Luca Gioffré, Luca Moroni, and Alberte Fernández-Castro gratefully acknowledge the support of AI Factory IT4LIA project and the CINECA support for access to high-performance computing facilities. Elena Marafatto

[7]https://policies.google.com/terms/generative-ai/use-policy

gratefully acknowledges the support of Agenzia per la Cybersicurezza Nazionale. Roberto Navigli acknowledges the support of PNRR MUR project PE0000013-FAIR.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Claude in order to: Grammar and spelling check, Formatting assistance, Improve writing style. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Su, Y. Lu, S. Pan, B. Wen, Y. Liu, Roformer: Enhanced transformer with rotary position embedding, ArXiv abs/2104.09864 (2021). URL: https://api.semanticscholar.org/CorpusID:233307138.

[2] B. Peng, J. Quesnelle, H. Fan, E. Shippole, YaRN: Efficient context window extension of large language models, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=wHBfxhZu1u.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI (2019). URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, accessed: 2024-11-15.

[4] A. Grattafiori, the Llama 3 Team, The Llama 3 Herd of Models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[5] A. Yang, B. Yu, C. Li, D. Liu, F. Huang, H. Huang, J. Jiang, J. Tu, J. Zhang, J. Zhou, J. Lin, K. Dang, K. Yang, L. Yu, M. Li, M. Sun, Q. Zhu, R. Men, T. He, W. Xu, W. Yin, W. Yu, X. Qiu, X. Ren, X. Yang, Y. Li, Z. Xu, Z.-Y. Zhang, Qwen2.5-1m technical report, ArXiv abs/2501.15383 (2025). URL: https://api.semanticscholar.org/CorpusID:275921951.

[6] G. Team, Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL: https://arxiv.org/abs/2403.05530. arXiv:2403.05530.

[7] G. Kamradt, Needle in a haystack - pressure testing llms, 2023. URL: https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

[8] Y. Kuratov, A. Bulatov, P. Anokhin, D. Sorokin, A. Sorokin, M. Burtsev, In search of needles in a 10m haystack: Recurrent memory finds what llms miss, 2024. arXiv:2402.10790.

[9] Y. Kuratov, A. Bulatov, P. Anokhin, I. Rodkin, D. Sorokin, A. Sorokin, M. Burtsev, Babilong: Testing the limits of llms with long context reasoning-in-a-haystack, in: A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems, volume 37, Curran Associates, Inc., 2024, pp. 106519–106554. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/c0d62e70dbc659cc9bd44cbcf1cb652f-Paper-Datasets_and_Benchmarks_Track.pdf.

[10] X. Zhang, Y. Chen, S. Hu, Z. Xu, J. Chen, M. K. Hao, X. Han, Z. L. Thai, S. Wang, Z. Liu, M. Sun, ∞bench: Extending long context evaluation beyond 100k tokens, 2024. URL: https://arxiv.org/abs/2402.13718. arXiv:2402.13718.

[11] C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, B. Ginsburg, RULER: What's the Real Context Size of Your Long-Context Language Models?, in: First Conference on Language Modeling, 2024. URL: https://openreview.net/forum?id=kIoBbc76Sy.

[12] F. Tamburini, BABILong-ITA: A new benchmark for testing large language models effective context length and a context extension method, in: C. Bosco, E. Jezek, M. Polignano, M. Sanguinetti (Eds.), Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, Cagliari, Italy, 2025, pp. 1112–1120. URL: https://aclanthology.org/2025.clicit-1.105/.

[13] L. Moroni, S. Conia, F. Martelli, R. Navigli, Towards a more comprehensive evaluation for Italian LLMs, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 584–599. URL: https://aclanthology.org/2024.clicit-1.67/.

[14] B. Magnini, M. Madeddu, M. Resta, R. Zanoli, M. Cimmino, P. Albano, V. Patti, A leaderboard for benchmarking LLMs on Italian, in: C. Bosco, E. Jezek, M. Polignano, M. Sanguinetti (Eds.), Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, Cagliari, Italy, 2025, pp. 636–646. URL: https://aclanthology.org/2025.clicit-1.61/.

[15] G. Puccetti, M. Cassese, A. Esuli, The invalsi benchmarks: measuring the linguistic and mathematical understanding of large language models in Italian, in: O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, S. Schockaert (Eds.), Proceedings of the 31st International Conference on Computational

Linguistics, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 6782–6797. URL: https://aclanthology.org/2025.coling-main.453/.

[16] M. Nissim, D. Croce, V. Patti, P. Basile, G. Attanasio, E. Musacchio, M. Rinaldi, F. Borazio, M. Francis, J. Gili, D. Scalena, B. Altuna, E. Azurmendi, V. Basile, L. Bentivogli, A. Bisazza, M. Bolognesi, D. Brunato, T. Caselli, S. Casola, M. Cassese, M. Cettolo, C. Collacciani, L. D. Cosmo, M. P. D. Buono, A. Esuli, J. Etxaniz, C. Ferrando, A. Fidelangeli, S. Frenda, A. Fusco, M. Gaido, A. Galassi, F. Galli, L. Giordano, M. Goffetti, I. Gonzalez-Dios, L. Gregori, G. Grundler, S. Iannaccone, C. Jiang, M. L. Quatra, F. Lagioia, S. M. Lo, M. Madeddu, B. Magnini, R. Manna, F. Mercorio, P. Merlo, A. Muti, V. Nastase, M. Negri, D. Onorati, E. Palmieri, S. Papi, L. Passaro, G. Pensa, A. Piergentili, D. Potertì, G. Puccetti, F. Ranaldi, L. Ranaldi, A. A. Ravelli, M. Rosola, E. S. Ruzzetti, G. Samo, A. Santilli, P. Santin, G. Sarti, G. Sartor, B. Savoldi, A. Serino, A. Seveso, L. Siciliani, P. Torroni, R. Varvara, A. Zaninello, A. Zanollo, F. M. Zanzotto, K. Zeinalipour, A. Zugarini, Challenging the abilities of large language models in italian: a community initiative, 2025. URL: https://arxiv.org/abs/2512.04759. arXiv:2512.04759.

[17] L. Moroni, G. Pappacoda, E. Barba, S. Conia, A. Galassi, B. Magnini, R. Navigli, P. Torroni, R. Zanoli, Sustainable Italian LLM evaluation: Community perspectives and methodological guidelines, in: C. Bosco, E. Jezek, M. Polignano, M. Sanguinetti (Eds.), Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, Cagliari, Italy, 2025, pp. 747–759. URL: https://aclanthology.org/2025.clicit-1.71/.

[18] B. Leite, T. F. Osório, H. L. Cardoso, FairytaleQA Translated: Enabling Educational Question and Answer Generation in Less-Resourced Languages, Springer Nature Switzerland, 2024, p. 222–236. URL: http://dx.doi.org/10.1007/978-3-031-72315-5_16. doi:10.1007/978-3-031-72315-5_16.

[19] Y. Xu, D. Wang, M. Yu, D. Ritchie, B. Yao, T. Wu, Z. Zhang, T. J.-J. Li, N. Bradford, B. Sun, T. B. Hoang, Y. Sang, Y. Hou, X. Ma, D. Yang, N. Peng, Z. Yu, M. Warschauer, Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 447–460. URL: https://aclanthology.org/2022.acl-long.34/. doi:10.18653/v1/2022.acl-long.34.

[20] L. Moroni, T. Bonomo, L. Gioffré, L. Xu, D. Fedele, L. Colosi, A. S. Bejgu, A. Scirè, R. Navigli, What we learned from continually training minerva: A case study on Italian, in: C. Bosco, E. Jezek, M. Polignano, M. Sanguinetti (Eds.), Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), CEUR Workshop Proceedings, Cagliari, Italy, 2025, pp. 760–774. URL: https://aclanthology.org/2025.clicit-1.72/.

[21] T. Bonomo, L. Gioffré, R. Navigli, LiteraryQA: Towards effective evaluation of long-document narrative QA, 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. 34086–34107. URL: https://aclanthology.org/2025.emnlp-main.1729/. doi:10.18653/v1/2025.emnlp-main.1729.

[22] S. Tedeschi, J. Bos, T. Declerck, J. Hajič, D. Hershcovich, E. Hovy, A. Koller, S. Krek, S. Schockaert, R. Sennrich, E. Shutova, R. Navigli, What's the meaning of superhuman performance in today's NLU?, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12471–12491. URL: https://aclanthology.org/2023.acl-long.697/. doi:10.18653/v1/2023.acl-long.697.

[23] F. M. Molfese, L. Moroni, L. Gioffré, A. Scirè, S. Conia, R. Navigli, Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 18477–18494. URL: https://aclanthology.org/2025.findings-acl.950/. doi:10.18653/v1/2025.findings-acl.950.

[24] M. Joshi, E. Choi, D. Weld, L. Zettlemoyer, TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1601–1611. URL: https://aclanthology.org/P17-1147/. doi:10.18653/v1/P17-1147.

[25] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A dataset for diverse, explainable multi-hop question answering, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2369–2380. URL: https://aclanthology.org/D18-1259/. doi:10.18653/v1/D18-1259.

[26] T. Kočiský, J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, E. Grefenstette, The narrativeqa

reading comprehension challenge, 2017. URL: https://arxiv.org/abs/1712.07040. arXiv:1712.07040.
[27] G. Comanici, the Gemini Team, Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL: https://arxiv.org/abs/2507.06261. arXiv:2507.06261.
[28] S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, in: J. Goldstein, A. Lavie, C.-Y. Lin, C. Voss (Eds.), Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909/.
[29] A. Lavie, A. Agarwal, METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments, in: C. Callison-Burch, P. Koehn, C. S. Fordyce, C. Monz (Eds.), Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 228–231. URL: https://aclanthology.org/W07-0734/.
[30] X. Wang, B. Ma, C. Hu, L. Weber-Genzel, P. Röttger, F. Kreuter, D. Hovy, B. Plank, "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7407–7416. URL: https://aclanthology.org/2024.findings-acl.441/. doi:10.18653/v1/2024.findings-acl.441.
[31] M. Polignano, P. Basile, G. Semeraro, Advanced natural-based interaction for the italian language: Llamantino-3-anita, 2024. URL: https://arxiv.org/abs/2405.07101. arXiv:2405.07101.
[32] R. Orlando, L. Moroni, P.-L. Huguet Cabot, S. Conia, E. Barba, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The first family of large language models trained from scratch on Italian data, in: F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024), CEUR Workshop Proceedings, Pisa, Italy, 2024, pp. 707–719. URL: https://aclanthology.org/2024.clicit-1.77/.
[33] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL: https://arxiv.org/abs/2306.05685. arXiv:2306.05685.

A. Prompts used

To ensure continuity with the QA items in the Summary-level split, we used essentially the same prompt used in [20], slightly tweaked, for *Local* and *Local alternative* items. For the Summary-level data split (the original INDAQA), the **summary** is provided as context and the model is asked to produce 20 items. Instead, for the Passage-level data split, we provide a **single passage** and ask to produce 3 QA items for *Local* and 1 for *Local (alternative)* setting (Figures 5 and 6, respectively). For *Entity* QA items, we devise a new prompt (Figure 7).

We report the prompt used for multiple-choice conversion (Figure 8) and the QA refinement step (in Figure 9).

```
User prompt
{text}
Domanda: {question}
{choices_block}
Risposta:
```

Figure 4: Prompt used for inference. In the generative task, the `choices_block` parameter is void.

B. Metadata Extraction

We extract bibliographic metadata for each document in our corpus by querying an LLM, using the Wikipedia pages of the documents to ground the generation process. We do not use Wikipedia metadata directly, as we found that for many pages they were missing or not updated with the information in the text. This appendix describes our metadata extraction pipeline and quality assurance procedures.

System prompt
```
Sei un esperto di letteratura. Il tuo compito è quello di generare domande e
risposte sulla trama di un testo letterario.
```

User prompt
```
TESTO: {text}

Scrivi almeno 3 domande diverse relative alla trama del testo. Per ogni domanda,
scrivi due possibili risposte, entrambe corrette e complete.
Le domande devono essere chiare e non ambigue. Se il testo è breve, genera almeno 2
domande.
Le risposte devono essere brevi e rispecchiare fedelmente il testo originale. Le
risposte possono anche essere quasi identiche.
Segui questo formato senza commentare:

Domanda: <domanda>
Risposta A: <risposta>
Risposta B: <risposta>
```

Assistant prompt
```
Domanda:
```

Figure 5: Prompt used to generate the *Local* QA items.

B.1. Extraction Pipeline

For each document in the corpus, we use the opening paragraph from its corresponding Wikipedia page as the primary information source. This introductory text typically contains the essential bibliographic information in a standardized format. We employ an LLM[8] to extract three metadata fields through separate, targeted prompts:

- **Publication year**: The year of first publication or composition
- **Author name**: The primary author or creator of the work
- **Title**: The canonical title of the work

Each field is queried independently to maximize extraction accuracy and allow for field-specific error handling. We show an example for extracting the publication year in Figure 10. Similar structured prompts are used for author names, titles, and genres, with appropriate formatting constraints specified for each field type.

B.2. Validation and Quality Control

To ensure metadata quality, we cross-reference the information with Wikipedia data. When structured metadata fields are available directly from Wikipedia, we compare the LLM-extracted values against these canonical sources. Discrepancies trigger manual review. When Wikipedia data are unavailable or ambiguous, we perform manual verification and correction.

C. Additional statistics

We report the list of the authorship distribution in Table 6, divided by data split. In the Summary-level split, a few authors dominate the corpus: *Carlo Goldoni* alone accounts for over one-third of the books (35.2%), followed by *Luigi Pirandello* (23.8%) and *Emilio Salgari* (13.3%). The remaining authors each contribute only a small fraction of the data, highlighting a strong skew toward a handful of prolific writers. This imbalance largely reflects the uneven coverage of authors on Wikipedia, as only documents for which a summary was available were included in the dataset.

[8] We found `meta-Llama/Llama-3.1-8B-Instruct` capable enough for this simple task.

System prompt

```
Sei un esperto di letteratura. Il tuo compito è quello di generare domande e
risposte sulla trama di un testo letterario.
```

User prompt

```
TESTO: {text}

Ecco degli esempi di domande e risposte riguardo questo testo:
{question_block}

Scrivi almeno un'altra domanda (diversa) relativa alla trama del testo. Per ogni
domanda, scrivi due possibili risposte, entrambe corrette e complete.
Le domande devono essere chiare e non ambigue. Se il testo è breve, genera comunque
almeno una domanda.
Le risposte devono essere brevi e rispecchiare fedelmente il testo originale. Le
risposte possono anche essere quasi identiche.
Segui questo formato senza commentare:

Domanda: <domanda>
Risposta A: <risposta>
Risposta B: <risposta>
```

Assistant prompt

```
Domanda:
```

Figure 6: Prompt used to generate the *Local (alternative)* QA items.

In the *Passage-level* split, the distribution is more balanced: the top author, *Enrico Castelnuovo*, represents 8.1% of the books, and most other authors contribute between 2% and 7%. This indicates that the passage-level split is more diverse with respect to authorship, and less dominated by a few prolific figures, complementing the summary-level data.

We also show the distribution of publication years for the two data splits in Figure 12. The Summary-level split spans a wide temporal range, with a mean publication year around 1814 and a large standard deviation (206 years), reflecting the broad historical coverage of the original material (strong outliers are listed in Table 7). In contrast, the Passage-level distribution is more homogeneous, with a mean publication year around 1900, a small standard deviation (21 years), and no strong outliers.

Finally, we present how the books are divided into 20 equal-probability bins based on text length quantiles and the chosen book for each bin in Figure 11.

D. Dataset examples

In this Section we present four examples taken from INDAQA2, in order to clarify the structure of the dataset.

E. Regexes

In Table 8, we report the regular expressions used to evaluate models in the multiple-choice setting. The list of regexes was defined by inspecting the outputs of selected models and identifying the most common patterns used to introduce the chosen option.

```
System prompt
Sei un esperto di letteratura. Dati degli estratti e delle domande e risposte
riguardanti un'entità (es., un personaggio letterario, un luogo, etc.), scrivi delle
nuove domande e risposte che siano più generali (meno specifiche riguardo a dettagli
del libro) ma sempre corrette e coerenti.

User prompt
Linee guida per le domande e risposte:
- Le domande devono essere chiare e pertinenti.
- Le domande e le risposte devono essere formulate in italiano corretto.
- Le domande, più che chiedere dettagli specifici, devono riguardare l'arco
narrativo del personaggio/il ruolo dell'entità nella storia.
---
{context_block}
---
Task:
Scrivi due/tre nuove domande riguardo {entity}, ognuna con una risposta, che siano
più generali ma sempre corrette e coerenti secondo le linee guida.
Segui esattamente questo formato senza commentare:
Domanda: <domanda>
Risposta: <risposta>

Assistant prompt
Domanda:
```

Figure 7: Prompt used to generate Entity QA items. The model is fed a list of passages and already generated QA items about the `entity` through `context_block`.

```
System prompt
Sei un esperto di letteratura. Dato un estratto, una domanda e le risposte giuste
che lo riguardano, genera tre "distrattori" (e.g., risposte sbagliate) plausibili.

User prompt
Libro: {title}
Estratto:
{context}
---
Domanda: {question}
Risposta A: {references[0]}
Risposta B: {references[1]}
---
Task: scrvi tre risposte sbagliate ma plausibili che potrebbero confondere un
lettore, basandoti sull'estratto fornito. Segui esattamente questo formato senza
commentare:
Distrattore X: <distrattore_x>

Assistant prompt
Distrattore 1:
```

Figure 8: Prompt used to generate Entity Question items starting from passages and existing questions.

```
System prompt
Sei un esperto di letteratura. Dato un estratto, decidi se la domanda e le risposte
che lo riguardano sono appropriate secondo le linee guida.

User prompt
Libro: {title}
Estratto:
{context}
---
Domanda: {question}
Risposta A: {references[0]}
Risposta B: {references[1]}
---
Linee guida per le domande e risposte:
- La domanda deve contenere abbastanza riferimenti in modo da non essere ambigua:
eventi e personaggi VANNO SPECIFICATI BENE con riferimenti temporali, col nome
proprio o con caratteristiche descritte nel testo, **in modo da renderla chiara
rispetto all'intero libro** (evita termini generici come 'un uomo', 'una donna',
'uno scontro', 'un incontro', etc.).
- La domanda deve essere formulata in modo da richiedere una risposta specifica e
non generica.
- La domanda non deve contenere frasi come 'nell'estratto fornito', 'nel libro',
'secondo te', 'cosa pensi di', 'come ti sembra', etc.; deve essere diretta e
neutrale.
- Le risposte devono essere ENTRAMBE corrette e complete; possono cambiare solo nel
modo in cui sono scritte (parafrasi).
- Le risposte non devono essere contenute nella domanda stessa.
- La domanda e le risposte devono essere formulate in italiano corretto.
---
Task:
Valuta se la domanda e le risposte sono appropriate secondo le linee guida. Se sono
appropriate, rispondi semplicemente con 'OK'. Se non lo sono, fornisci una versione
corretta della domanda e delle risposte seguendo le linee guida ed usando l'estratto
fornito; riscrivi tutti gli elementi corretti.
```

Figure 9: Prompt used to correct the generated QA items.

```
User prompt
Based on the following text from a Wikipedia page, extract only the publication year
of the work. Respond with just the year as a four-digit number.
{wikipedia_paragraph}

Assistant prompt
Year:
```

Figure 10: Prompt used to extract the publication year from the starting paragraph of a Wikipedia page.

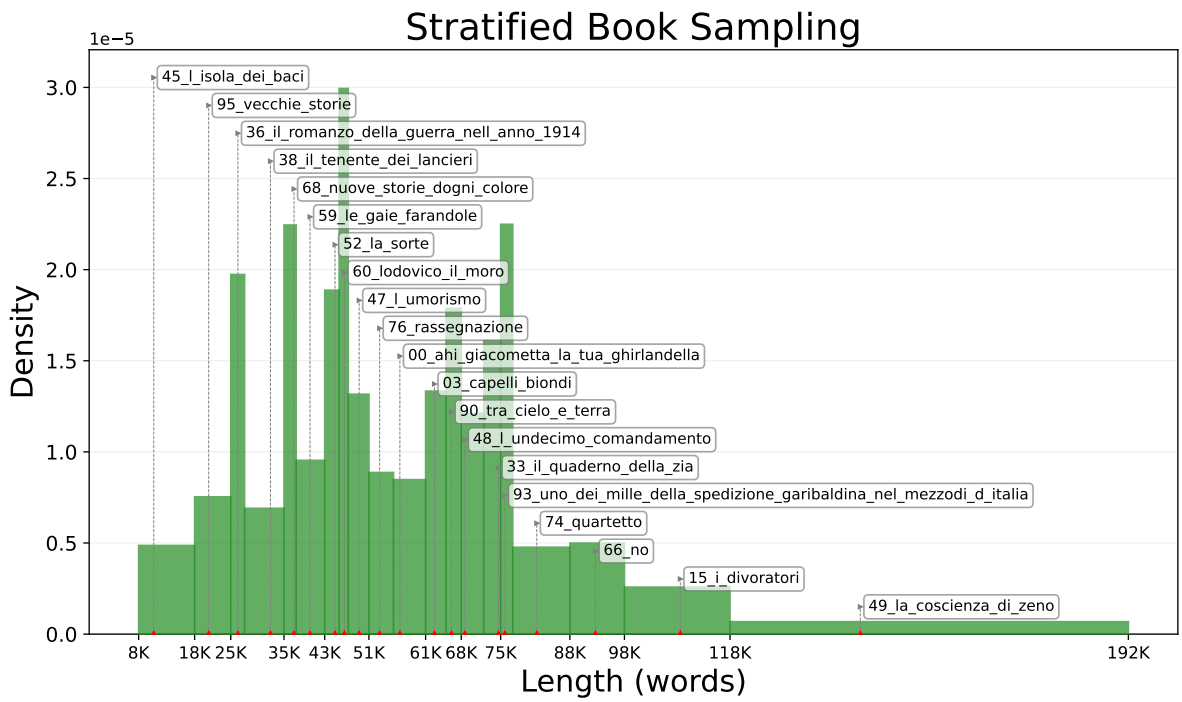

Figure 11: Distribution of the books in the annotation set. We show the ID of the selected 20 books, one from each of the 20 equal-probability bins.

Summary-Level			*Passage-Level*		
Author	**# Books**	**%**	**Author**	**# Books**	**%**
Carlo Goldoni	127	35.2	Enrico Castelnuovo	8	8.1
Luigi Pirandello	86	23.8	Matilde Serao	7	7.1
Emilio Salgari	48	13.3	Antonio Beltramelli	5	5.1
Edgar Allan Poe	19	5.3	Edmondo De Amicis	5	5.1
Grazia Deledda	10	2.8	Guido da Verona	5	5.1
Giuseppe Giacosa	7	1.9	Alfredo Panzini	5	5.1
Pietro Metastasio	6	1.7	Luigi Capuana	4	4.0
Jules Verne	4	1.1	Anton Giulio Barrili	4	4.0
William Shakespeare	3	0.8	Alfredo Oriani	3	3.0
Antonio Fogazzaro	3	0.8	Nicola Misasi	3	3.0
Niccolò Machiavelli	2	0.6	Emilio De Marchi	3	3.0
Charles Perrault	2	0.6	Annie Vivanti	3	3.0
Giovanni Verga	2	0.6	F.T. Marinetti	3	3.0
Paolo Mantegazza	2	0.6	Salvatore Farina	2	2.0
Matilde Serao	2	0.6	Lucio D'Ambra	2	2.0
Giovanni Pascoli	2	0.6	Jack London	2	2.0
Alessandro Verri	2	0.6	Gerolamo Rovetta	2	2.0
			Federico De Roberto	2	2.0
			E.A. Butti	2	2.0
			Luigi Pirandello	2	2.0
			Italo Svevo	2	2.0

Table 6
Authors with more than one book in the benchmark, divided by *Summary-level* and *Passage-level* splits, along with the number of books and relative percentages.

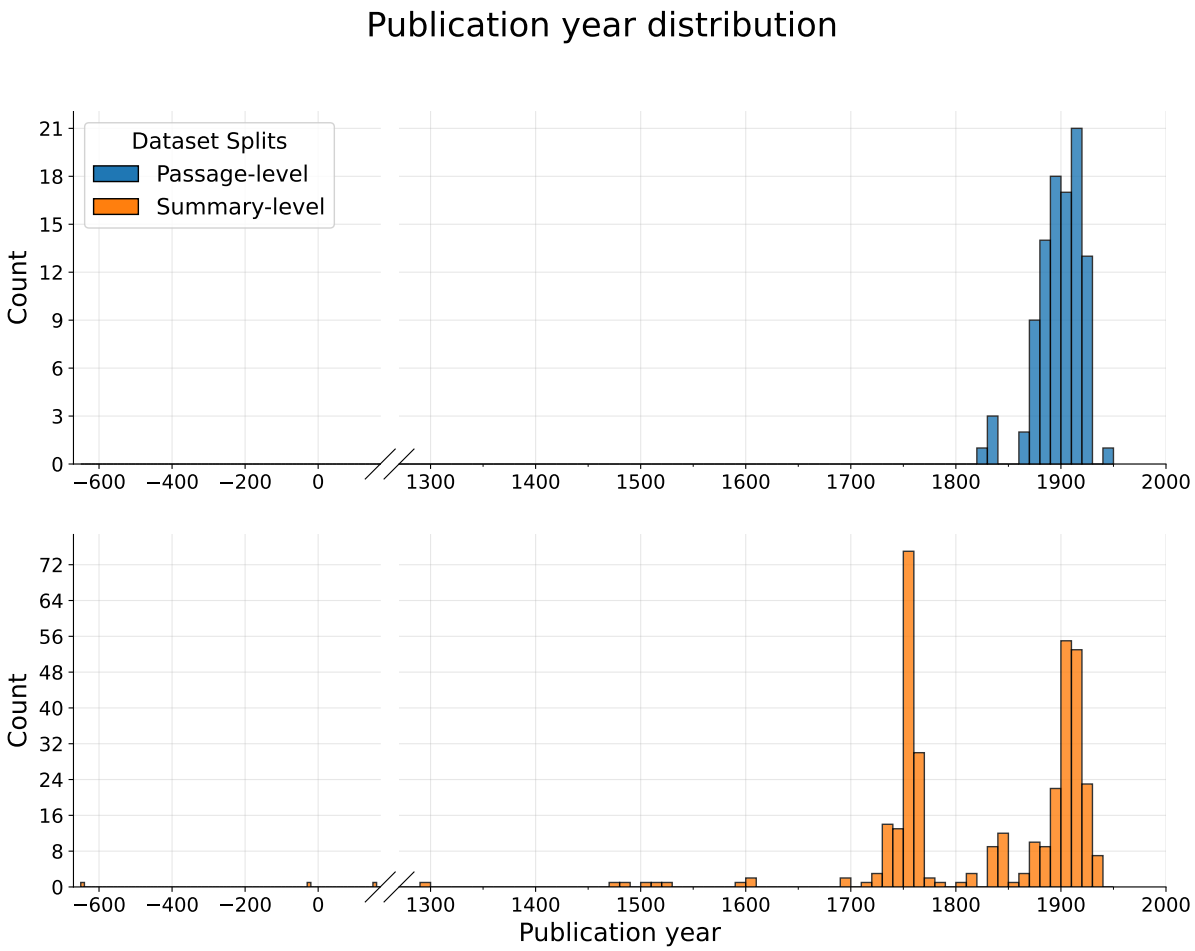

Figure 12: Publication year distribution of the books in the benchmark. Years have been extracted from the introductory paragraph of the related Wikipedia page (either automatically or manually).

Title	Author	Publication year
Lo scudo di Eracle	Esiodo	~650 BCE
Satire	Quinto Orazio Flacco	~30 BCE
La storia vera	Luciano di Samosata	~150 CE
Il Milione	Marco Polo	1298 CE
Stanze per la giostra	Angelo Poliziano	1475 CE
Orlando innamorato	Matteo Maria Boiardo	1483 CE
Arcadia	Jacopo Sannazaro	1504 CE
Mandragola	Niccolò Machiavelli	1518 CE
Clizia	Niccolò Machiavelli	1525 CE

Table 7
Books from the *Summary-level* data split that are outliers with respect to the publication year distribution.

```
// summary_level
{
    "answers": [
        "In un villaggio della Foresta Nera.",
        "Nella Foresta Nera, in un villaggio."
    ],
    "choices": [],                  // not available
    "entity": null,                 // not available
    "kind": "summary_question",
    "model": "gemini2-flash",
    "question": "Dove si svolge la festa di fidanzamento iniziale?",
    "question_id": "000_le_villi.summary.0",
    "source_paragraphs_ids": [],    // not available
    "source_questions_ids": [],     // not available
    "target": {
        "label": null,              // not available
        "text": null                // not available
    }
}
```

Figure 13: QA item from the summary-level data split. Please note the null/empty data fields for this combination.

```
// passage_level - Local Question item
{
    "answers": [
        "Giacometta Maldi",
        "Giacometta"
    ],
    "choices": [
        "A. Carolina",
        "B. Elena",
        "C. Giacometta Maldi",
        "D. Geltrude"
    ],
    "entity": null,                     // not available
    "kind": "local_question",
    "model": "gemini-2.5-flash",
    "question": "Come si chiama la giovane donna al centro delle attenzioni per il
     ↪  matrimonio?",
    "question_id": "00_ahi_giacometta_la_tua_ghirlandella.set-a.1",
    "source_paragraphs_ids": [0],
    "source_questions_ids": [],         // not available
    "target": {
        "label": "C",
        "text": "Giacometta Maldi"
    }
}
```

Figure 14: QA item from the Passage-level data split, Local Question set. Please note the null/empty data fields for this combination.

```
// passage_level - Alternative Local Question item
{
    "answers": [
        "Biondi",
        "Erano biondi"
    ],
    "choices": [
        "A. Neri",
        "B. Biondi",
        "C. Castani",
        "D. Rossi"
    ],
    "entity": null,                     // not available
    "kind": "local_question_alt",
    "model": "gemini-2.5-flash",
    "question": "Di che colore erano i capelli di Giacometta?",
    "question_id": "00_ahi_giacometta_la_tua_ghirlandella.set-b.1",
    "source_paragraphs_ids": [0],
    "source_questions_ids": [],         // not available
    "target": {
        "label": "B",
        "text": "Biondi"
    }
}
```

Figure 15: QA item from the Passage-level data split, Local Alternative Question set. Please note the null/empty data fields for this combination.

```
// passage_level - Entity Questions
{
    "answers": ["La sua eccentricità e la tendenza a comportarsi in modo inappropriato o
     ↪  fuori luogo."],
    "choices": [],                      // not available
    "entity": "adalgisa",
    "kind": "entity_question",
    "model": "gemini-2.5-flash",
    "question": "Qual è una caratteristica distintiva del personaggio di Adalgisa?",
    "question_id": "00_ahi_giacometta_la_tua_ghirlandella.",
    "source_paragraphs_ids": [4, 8],
    "source_questions_ids": [0, 2, 4],
    "target": {
        "label": null,                  // not available
        "text": null                    // not available
    }
}
```

Figure 16: QA item from the Passage-level data split, Entity Question set. Please note the null/empty data fields for this combination.

#	Regex
1	`r"Risposta \(?([ABCD])\)?"`
2	`r"risposta corretta è la \(?([ABCD])\)?"`
3	`r"risposta corretta è: \(?([ABCD])\)?"`
4	`r"risposta corretta è \(?([ABCD])\)?"`
5	`r"risposta \(?([ABCD])\)?"`
6	`r"Risposta: \(?([ABCD])\)?"`
7	`r"risposta: \(?([ABCD])\)?"`
8	`r"Risposta è \(?([ABCD])\)?"`
9	`r"risposta è \(?([ABCD])\)?"`
10	`r"Risposta è la \(?([ABCD])\)?"`
11	`r"risposta è la \(?([ABCD])\)?"`
12	`r"\n\(?([ABCD])\)?\. "`
13	`r"\n\(?([ABCD])\)? "`
14	`r"\(?([ABCD])\)?\.? "`
15	`r"^([ABCD])\.?$"`

Table 8
Regex patterns used to extract the answer from the models' responses in the MCQA task.