PharmaQA.IT: an Italian dataset for Q&A in the pharmaceutical domain

Kamyar Zeinalipour
University of Siena / Siena, Italy
Yukai / Siena, Italy
kamyar.zeinalipour2@unisi.it

Andrea Zugarini
expert.ai / Siena, Italy
azugarini@expert.ai

Asya Zanollo
Istituto Universitario di Studi Superiori / Pavia, Italy
asya.zanollo@iusspavia.it

Leonardo Rigutini
expert.ai / Siena, Italy
lrigutini@expert.ai

Abstract

The growing use of Large Language Models (LLMs) for medical Question Answering (QA) requires reliable, evidence-grounded benchmarks beyond English. In Italy, Riassunti delle Caratteristiche del Prodotto (RCP) issued by the Italian Medicines Agency (AIFA) are the main regulatory source on medicines, yet no QA dataset exists on these documents, limiting the development and evaluation of trustworthy Italian QA systems.

We introduce **PharmaQA.IT**, an Italian extractive QA dataset built from RCPs in PharmaER.IT. Using a semi-automatic pipeline, we (i) select informative pages from 1,077 leaflets, (ii) prompt a multimodal LLM on page images with professional personas to generate candidate question–answer pairs, and (iii) validate and normalise them with expert revision. The final dataset contains 861 high-quality question–answer pairs on indications, contraindications, dosage, warnings, interactions, and pharmacological properties.

We frame PharmaQA.IT as an extractive QA benchmark with structured JSON outputs and evaluate a range of open and proprietary LLMs. Results show that open models approach closed-source performance under a chunking-and-retrieval setup. PharmaQA.IT, together with all code, prompts, and evaluation scripts, will be publicly released to support research on trustworthy Italian biomedical QA. PharmaQA.IT, together with all code, prompts, and evaluation scripts, is publicly available on Hugging Face to support research on trustworthy Italian biomedical QA.

1 Introduction

The growing use of Large Language Models for medical Question Answering (QA) increases the need for reliable, evidence-grounded benchmarks beyond English. Existing medical QA datasets are mostly English and centred on scientific articles or clinical notes (Tsatsaronis et al., 2015b; Pampari et al., 2018; Ben Abacha et al., 2019), while multilingual resources target general-domain content (Artetxe et al., 2020b; Lewis et al., 2020a; Clark et al., 2020a). For Italian, most NLP datasets focus on general-domain NER and syntax (Bosco, 2000; Magnini et al., 2006; Basile et al., 2012, 2016, 2020; Tedeschi and Navigli, 2022); in the pharmaceutical domain, PharmaER.IT (Zugarini and Rigutini, 2025b) covers medical NER over *Riassunti delle Caratteristiche del Prodotto* (RCP), but no QA benchmark exists on these regulatory documents.

Example instance from PHARMAQA.IT

Context (RCP excerpt).
Il Riassunto delle Caratteristiche del Prodotto per la soluzione di glucosio 5% (sacca Viaflo) descrive il periodo di validità delle diverse confezioni (50–1000 ml) quando il medicinale è conservato non aperto. In particolare, per la sacca da 1000 ml viene indicata una durata di conservazione di 3 anni.

Question.
Qual è il periodo di validità della sacca da 1000 ml di Glucosio 5% non aperta?

Extractive answer.
3 anni

Figure 1: Illustrative question–answer pair in PHARMAQA.IT derived from an Italian Summary of Product Characteristics (RCP).

We address extractive QA over Italian RCPs issued by AIFA: given a question and the corresponding leaflet, a system must return a concise answer strictly grounded in the document, with explicit evidence spans and no hallucination. We introduce **PharmaQA.IT**, built from 1,077 RCPs in PharmaER.IT via a semi-automatic pipeline that selects informative pages, prompts a multimodal LLM with professional personas to propose question–answer pairs, and then validates and normalises them with expert revision. The final dataset contains 861 high-quality pairs on indications, con-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 937–947
March 25-27, 2026

traindications, dosage, warnings, interactions and pharmacological properties, each linked to the full source RCP and represented in a structured JSON format with evidence.

Our study is guided by three questions: (RQ1) how difficult is extractive QA over Italian RCPs for current LLMs; (RQ2) how baseline retrieval and chunking strategies affect answer accuracy and evidence selection; and (RQ3) how close strong open models can get to proprietary APIs under the same constrained setting. We evaluate a broad set of open- and closed-source LLMs employing a standard retrieval-augmented setup as a baseline, varying chunk sizes and overlaps to expose the impact of document segmentation on performance.

Figure 1 shows an example question–answer pair from PHARMAQA.IT. Our contributions are threefold: we release the first Italian QA benchmark on pharmaceutical regulatory documents (PharmaQA.IT); we describe a reusable semi-automatic pipeline for deriving QA data from long, noisy RCP PDFs using multimodal LLMs plus expert validation; and we provide an extensive comparison of open and proprietary LLMs using a canonical RAG baseline, showing that competitive open models can approach closed-source systems, when retrieval and chunking are carefully designed.

2 Related works

Specialized medical Question Answering (QA) is well-established in English via benchmarks like BioASQ (Tsatsaronis et al., 2015a) and emrQA (Pampari et al., 2018), with challenges such as MEDIQA (Ben Abacha et al., 2019) expanding evaluation to include inference and entailment. While high-quality multilingual resources exist—notably XQuAD (Artetxe et al., 2020a), MLQA (Lewis et al., 2020a), and TyDiQA (Clark et al., 2020b)—they predominantly target general domains rather than clinical documentation. In Italian, resources remain scarce; although PharmaER.IT (Zugarini and Rigutini, 2025a) recently addressed Named Entity Recognition (NER) on "Riassunti delle Caratteristiche del Prodotto" (RCP), no dedicated QA benchmark currently exists for these critical regulatory artifacts. Early initiatives focused on large-scale biomedical QA benchmarks, such as BioASQ (Tsatsaronis et al., 2015b), which provides factoid and list-based questions from PubMed, or emrQA (Pampari et al., 2018), derived from clinical notes in electronic health records. Other challenges, including MEDIQA (Ben Abacha et al., 2019), expanded evaluation to natural language inference and summarization across medical documents. In multilingual contexts, resources like XQuAD (Artetxe et al., 2020b), MLQA (Lewis et al., 2020a), and TyDiQA (Clark et al., 2020a) offer high-quality cross-lingual QA benchmarks but focus on general-domain content and do not include regulatory or pharmaceutical documents. Conversely, biomedical NLP has also progressed through domain-specific corpora for related tasks such as NER, relation extraction, and document classification. Examples include BioCreative (Li et al., 2016), the distant-supervision biomedical corpora of Quirk et al. (2016), and more recent automated or semi-automated annotation approaches (Menezes and Roth, 2019; Alves and Coheur, 2022; Zhou et al., 2023; Ringland et al., 2019). However, these resources remain predominantly English-centric.

Within the Italian landscape, most NLP datasets target general-domain NER (Bosco, 2000; Magnini et al., 2006; Basile et al., 2012, 2016, 2020) or Wikipedia-derived corpora such as MultiNERD (Tedeschi and Navigli, 2022). Italian datasets in specialized vertical domains are scarce.

Recently, (Zugarini and Rigutini, 2025b) introduced PharmaER.IT, a NER dataset in the pharmaceutical domain. It was made of annotated Italian drug leaflets for medical NER, providing both gold and silver annotations from AIFA regulatory documents. However, no QA-oriented dataset existed for this domain. PharmaQA.IT fills this gap by introducing the first Italian QA benchmark built from pharmaceutical regulatory documentation.

Retrieval-Augmented QA. Recently, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020b; Izacard and Grave, 2022) has emerged as a promising paradigm for knowledge-intensive QA, combining neural language models with document retrieval to improve factual accuracy and scalability. By integrating a retriever component with a generator, RAG allows models to access relevant external knowledge dynamically, reducing hallucinations and improving answer precision. While RAG and similar approaches have been widely applied in English biomedical and general-domain QA (Lewis et al., 2020b; Izacard and Grave, 2022; Guu et al., 2020), their adoption in Italian, particularly in pharmaceutical regulatory contexts, remains unexplored. PharmaQA.IT could serve as

a benchmark for investigating RAG-based Italian QA systems, enabling experiments that leverage both neural reasoning and retrieval over official drug documents.

3 The Dataset

PharmaQA.IT is constructed by re-purposing the pharmaceutical documents contained in the PharmaER.IT dataset and enriching them with high-quality question–answer pairs suitable for comprehension-oriented tasks.

3.1 Dataset Creation Methodology

The dataset was created through a semi-automated annotation pipeline designed to balance efficiency and quality. First, (i) a subset of RCPs from PharmaER.IT was selected to ensure coverage of different therapeutic classes. Next, (ii) a generation module proposed candidate question-answer pairs by extracting answer spans directly from the text. Finally, (iii) domain experts reviewed these automatically generated QA pairs, validating, correcting, or discarding them to ensure factual accuracy and appropriateness. Questions were designed to reflect realistic information needs of healthcare professionals and patients, covering topics such as adverse reactions, indications, administration guidelines, and drug–drug interactions.

Document selection. As the primary source of textual material, we employ PharmaER.IT (Zugarini and Rigutini, 2025b), a dataset developed for Named Entity Recognition (NER) in the Italian pharmaceutical domain. PharmaER.IT includes Riassunti delle Caratteristiche del Prodotto (RCP), the official regulatory documents for all drugs marketed in Italy, and is divided into two corpora: (i) Gold, with manually annotated and validated documents, and (ii) Silver, with automatically annotated documents lacking expert validation. To generate QA examples we used the Silver corpus for its size and greater variability. Each pharmaceutical leaflet was preprocessed by converting all PDF pages into PNG images. Only documents with at least 10 pages were retained, in order to ensure sufficient content heterogeneity. The final dataset was built from a collection of 1077 documents (after the filtering step), representing a broad array of pharmaceutical product types. Since pharmaceutical leaflets typically include mixed content — tables, plots/figures, and text-heavy sections — the dataset aims to reflect this diversity. For each document, pages were iteratively sampled in random order and passed to a multimodal LLM (Qwen3-VL-235B-A22B), explicitly selected for its robust optical character recognition (OCR) and document layout analysis capabilities. Processing the image, instead of the plain text, avoids incurring OCR issues and yields a full view of the document content, pictures, and layout included. Depending on the content of the page, the LLM establishes whether there are plots/figures, tables or plain text. For each category, at most one representative page per document was selected.

QA Pairs Generation To produce coherent question-answer pairs we ground the generation to the content of a page, following an approach similar to (Zugarini et al., 2024b,a). For each selected page, a synthetic question–answer pair was generated by prompting the multimodal model (Qwen3-VL-235B-A22B). Depending on the detected content type, a dedicated prompt was used. Moreover, for each prompt type, we defined "simple" and "complex" variants, the latter chosen with a 25% probability of selecting the complex form. This allowed to vary the complexity of the generated QA pairs and to control over linguistic and reasoning difficulty. The "simple" variants prompts are presented in Appendix A. To promote variability in language style and domain framing, each prompt was conditioned on a randomly sampled persona from a curated list of Italian pharmaceutical professional profiles (e.g., pharmacist, clinical researcher, regulatory specialist).

Human validation To assess the linguistic reliability of the generated QA pairs, human validation was performed by native Italian speakers with master's degrees in Linguistics and expertise in AI-generated data validation. We prioritized linguistic expertise over clinical background because the objective was to verify strict textual grounding and linguistic well-formedness (extractive verification), rather than to assess external medical validity. The generated questions are either wh-type open questions or yes/no questions, targeting the exact information present in the document. Questions also contain a reference to a context. The answer structure strongly depends on the question type. In general, answers tend to be concise, reporting the exact information from the document without rephrasing or elaboration. The evaluation assessed the following criteria: (i) Comprehensibility of both Question and Answer; (ii) Question quality

and naturalness; (iii) Answer completeness; (iv) Answer Relevance to the Question; (v) Reliability with respect to the provided reference. Evaluators used a binary rating scale (accept/discard) and performed independent annotation to ensure unbiased assessment. Examples for accepted and discarded QA pairs are provided in Appendix B.

3.2 Dataset Statistics

PharmaQA.IT contains **861** question–answer pairs, each linked to a full RCP. Table 1 summarizes length statistics in both words and subword tokens[1]. Questions have an average length of **29.8** tokens (**16.0** words), while answers are generally concise, averaging **17.8** tokens (**6.9** words). In contrast, the associated RCP documents are extensive, with a mean length of **15,149.6** tokens (**7,021.5** words). This highlights a strong length mismatch between the short queries and the long context required to answer them.

Figure 2 shows token-length distributions for questions, answers and full texts. All are right-skewed: most questions and answers are short with a long tail of longer cases, while documents cover a wide range but concentrate around 10k–20k tokens. This combination of short QA spans and long, heterogeneous RCPs makes PharmaQA.IT a realistic and challenging benchmark for regulatory QA.

4 Experiments

In this section we describe the experimental setup used to evaluate PharmaQA.IT as a benchmark for Italian-domain Question Answering in the pharmaceutical domain. We focus on an extractive QA setting in which Large Language Models (LLMs) must answer questions using only the content of the corresponding Riassunto delle Caratteristiche del Prodotto (RCP), and we systematically compare a wide range of open- and closed-source models, different document chunking strategies, and multiple automatic evaluation metrics.

4.1 Task Formulation

Each PharmaQA.IT instance consists of a question, a short answer span, and the associated RCP. Given the question and the RCP, the model must return a concise answer *strictly* grounded in the document, together with explicit evidence. We cast this as extractive QA with structured output: the model receives a "Context" made of one or more RCP chunks and, following an Italian system prompt,[2] must answer only from this Context and output a JSON object with `answer`, `evidence` (pairs of `chunk_id` and verbatim `quote`), `chunks_used`, and `status` (`ok`, `non_present`, or `ambiguous`). This setup forces models to act as extractive readers rather than general conversational agents and mirrors the structure of the gold annotations.

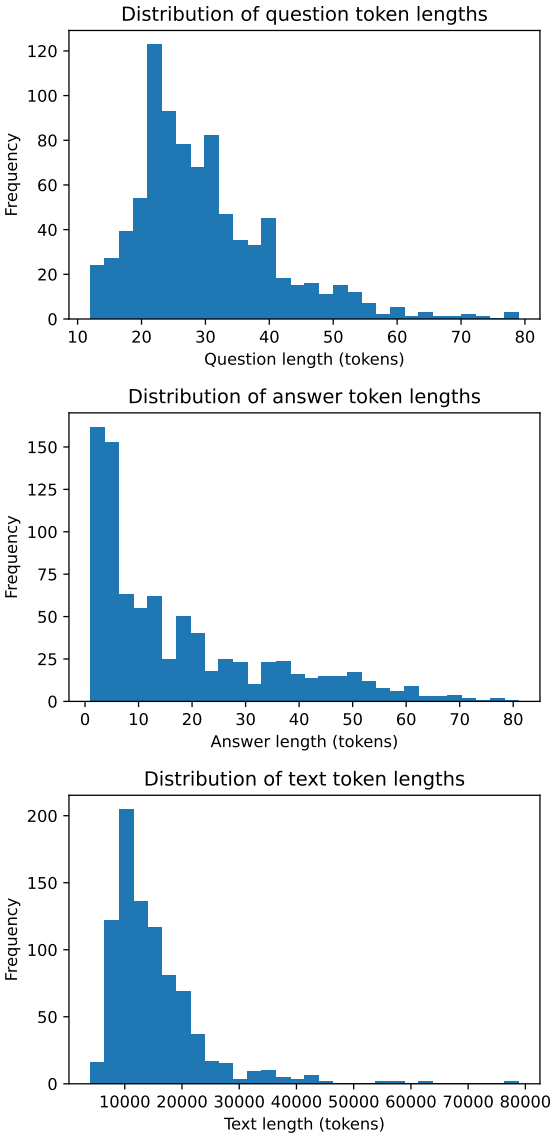

Figure 2: Distributions of token lengths for questions, answers and full RCP texts in PharmaQA.IT.

[1]We count tokens with the `meta-llama/Meta-Llama-3-8B-Instruct` tokenizer to ensure consistency with downstream experiments.

[2]The full prompt and JSON schema are given in Appendix C.

Span	Unit	Count	Min	Max	Mean	Median	Std. dev.
Question	Tokens	861	12	79	29.80	27.00	10.87
	Words	861	6	49	15.97	14.00	6.36
Answer	Tokens	861	1	81	17.82	11.00	17.11
	Words	861	1	38	6.92	4.00	6.91
Text	Tokens	861	3,992	78,809	15,149.57	13,153.00	8,394.84
	Words	861	1,673	36,008	7,021.51	6,042.00	3,893.57

Table 1: Length statistics in subword tokens (via `Meta-Llama-3-8B-Instruct`) and words for questions, answers, and full RCP texts in PharmaQA.IT.

4.2 Models

We evaluate a heterogeneous set of Instruction-tuned LLMs covering both open-source checkpoints (run locally) and proprietary models accessed via APIs.[3]

Open-source models. We evaluate 16 instruction-tuned LLMs from 1B to 30B parameters, covering both general-purpose and European/Italian-oriented models from the Llama 3, Salamandra, Mistral, DeepSeek, Gemma, SmolLM3, Apertus, EuroLLM, Minerva and Qwen3 families.

Closed-source API models. We further compare with five proprietary APIs, all queried with the same Italian system prompt and JSON output format: `gpt-4.1`, `deepseek-chat`, `qwen-plus`, `mistral-large-latest`, and `gemini-2.5-flash`.

4.3 Chunking and Context Construction

Since pharmaceutical RCPs are lengthy documents and model context windows are often limited, we segment each document into overlapping chunks (CHUNK_TOKENS, CHUNK_OVERLAP). This strategy ensures that relevant information is accessible to the model within its context window. We utilize the following settings:

$$[64, 16],\ [128, 32],\ [256, 64],\ [512, 128].$$

For each question we take the RCP containing the gold answer, split it accordingly, and assign each segment an explicit label `[Chunk N]` used in the `chunk_id` and `chunks_used` fields of the JSON output. Unless otherwise stated, we rank chunks with a dense retriever based on `multilingual-e5-base` embeddings (cosine similarity) and build the LLM "Context" by concatenating the top-k segments ($k = 10$) in document order.[4]

4.4 Prompting and Generation Settings

All models are prompted in Italian with the same system prompt, which defines the model as an *assistente estrattivo*, forbids using information outside the Context, and enforces a fixed JSON schema (see Appendix C). For open-source models we set `MAX_NEW_TOKENS=256`, `TEMP=0.0`, `TOPK=10`, `TOPP=0.95`; closed-source APIs use equivalent settings. A zero temperature makes generation effectively deterministic and facilitates comparison.

4.5 Evaluation Metrics

We use two automatic metrics on the `answer` field of the model JSON. **Exact Match (EM)** checks whether the prediction exactly matches the gold span after lowercasing and trimming, giving a strict measure of correctness. **ROUGE-L $\mathbf{F_1}$** computes longest-common-subsequence overlap between predicted and gold answers, allowing minor paraphrases and inflectional variation and providing a softer similarity score than EM.

The combination of EM and ROUGE-L jointly captures exact-span recovery and approximate textual similarity, which is crucial for semantic adequacy in a specialised, safety-critical setting like pharmaceutical QA.

4.6 Results and Discussion

Open-source models. Table 2 reports ROUGE-L F_1 all open-source models across the four chunk configurations. We explicitly limit our evaluation to efficient models in the `1B-30B` range (the 'edge' class) to test viability for local deployment in privacy-sensitive pharmaceutical environments, thereby excluding larger

[3] All open-source models are run on a server equipped with two NVIDIA RTX A6000 GPUs (48GB VRAM each). Closed-source models are accessed through remote APIs and do not require local GPU resources.

[4] We keep $k = 10$ fixed and only vary chunk size and overlap to isolate segmentation effects.

70B+ parameter checkpoints. Overall, both metrics improve with chunk size, from 64–16 to 512–128, for most strong models. Among Llama, Llama-3.1-8B-Inst reaches the best scores (F_1 0.618, EM 0.386) at 512–128; for Mistral, Mistral-Small-24B-Inst is the strongest open model (F_1 0.656, EM 0.433), followed by Mistral-7B-Inst. Qwen3-30B-A3B-Inst and Mistral-Nemo-Inst also peak at 512–128 (around F_1 0.61, EM 0.36).
In contrast, smaller or less aligned models perform poorly: Salamandra variants stay below F_1 0.10 in all settings, and some European-focused checkpoints (EuroLLM-9B-Inst, Minerva-7B-Inst) even degrade with larger chunks, suggesting difficulties with long contexts and strict JSON formatting. Overall, competitive open models clearly benefit from larger chunks, which expose more of the RCP and facilitate evidence aggregation and exact-span recovery, as reflected in consistent EM gains.

Retrieval analysis. Table 3 reports retrieval hit@10 for the four chunk configurations, averaged over open-source models. The hit rate increases with chunk size, from 0.436 at 64–16 to **0.615** at 512–128: larger chunks make it more likely that the gold span appears fully in at least one of the top-10 passages, despite fewer chunks per document. The best configuration (512–128) also yields the strongest QA scores (Table 2), so, to decouple chunking from model quality, we evaluate all closed-source models only under 512–128.

Closed-source models and comparison. Table 4 reports closed-source models, evaluated only with the 512–128 configuration. The best API, deepseek-chat, reaches F_1 **0.679** and EM **0.432**, followed by GPT-4.1 (F_1 0.636, EM 0.367) and Qwen-Plus (F_1 0.579, EM 0.328); Mistral-Large is weaker (F_1 0.477, EM 0.237), and Gemini-2.5-Flash lags further behind. Compared with open models in Table 2, the gap is modest: Mistral-Small-24B-Inst attains F_1 0.656 and EM 0.433, essentially matching deepseek-chat on EM while only a few ROUGE-L F_1 points behind; Llama-3.1-8B-Inst (F_1 0.618) and Qwen3-30B-A3B-Inst (F_1 0.609) rival Qwen-Plus and approach GPT-4.1. Thus, in PharmaQA.IT's constrained extractive setting with tuned chunking and retrieval, strong open models can approach or match closed systems. The benchmark remains challenging, with all closed models below F_1 0.68 and EM 0.43, underscoring the need for better retrieval and extraction and motivating our focus on 512–128.

5 Conclusion

We introduced **PharmaQA.IT**, the first Italian benchmark for extractive Question Answering over pharmaceutical regulatory documents. Starting from the RCPs in PharmaER.IT, we built a semi-automated pipeline that selects informative pages from 1,077 AIFA leaflets, uses a multimodal LLM with professional personas to propose question–answer pairs, and validates and normalises them through expert revision. The dataset contains 861 QA pairs on indications, dosage, contraindications, warnings, interactions and pharmacological properties, each linked to the full source RCP and represented in a JSON schema with explicit evidence spans.
Experiments with a broad set of open and closed LLMs, under different chunking and retrieval configurations, show that PharmaQA.IT is challenging (RQ1), that larger chunks which increase the chance of retrieving the full answer span substantially improve performance (RQ2), and that strong open-source models can approach, and sometimes match, proprietary APIs in exact-match accuracy under the same pipeline (RQ3). PharmaQA.IT thus serves both as a research resource and as a realistic benchmark for industrial stakeholders to compare QA engines and assess evidence tracking. Future work includes extending the dataset with more diverse question types (e.g., multi-hop, unanswerable and patient-oriented questions), exploring domain-adaptive training for Italian and multilingual LLMs, and moving towards multimodal QA over tables and figures in RCPs.

6 Limitations

While PharmaQA.IT provides the first Italian benchmark for extractive QA over pharmaceutical regulatory documents, it also comes with a number of limitations that should be taken into account when interpreting our results.

Domain and language coverage. PharmaQA.IT is restricted to Italian *Riassunti delle Caratteristiche del Prodotto* (RCP) issued by AIFA. As such, it does not cover other document types that are relevant in practice, such as patient information leaflets, clinical notes, guidelines, or scientific literature,

Model	64-16		128-32		256-64		512-128	
	F1	EM	F1	EM	F1	EM	F1	EM
Llama-3.2-1B	0.100	0.035	0.145	0.072	0.097	0.045	0.115	0.060
Llama-3.2-3B	0.413	0.204	0.466	0.230	0.525	0.272	0.564	0.314
Llama-3.1-8B	0.475	0.252	0.527	0.304	0.579	0.340	0.618	0.386
Salamandra-2B	0.036	0.001	0.041	0.000	0.038	0.000	0.036	0.000
Salamandra-7B	0.088	0.017	0.082	0.014	0.070	0.002	0.055	0.000
Mistral-7B	0.426	0.192	0.462	0.232	0.531	0.296	0.545	0.304
Mistral-Small-24B	0.417	0.253	0.492	0.312	0.581	0.379	**0.656**	**0.433**
Mistral-Nemo	0.456	0.236	0.525	0.294	0.597	0.350	0.610	0.361
DeepSeek-V2-Lite	0.154	0.017	0.137	0.005	0.169	0.023	0.174	0.035
DeepSeek-7B	0.298	0.118	0.222	0.055	0.173	0.007	0.005	0.000
Gemma-3-12B	0.369	0.122	0.412	0.131	0.441	0.143	0.480	0.150
SmolLM3-3B	0.381	0.185	0.406	0.204	0.472	0.251	0.449	0.230
Apertus-8B	0.479	0.256	0.488	0.269	0.526	0.302	0.546	0.325
EuroLLM-9B	0.366	0.166	0.338	0.130	0.235	0.030	0.020	0.000
Minerva-7B	0.220	0.000	0.203	0.001	0.177	0.000	0.033	0.001
Qwen3-30B	0.464	0.254	0.522	0.302	0.577	0.339	0.609	0.361

Table 2: Results on PharmaQA.IT for open-source models across different chunk sizes. We report ROUGE-L F1 (F1) and Exact Match (EM).

Chunk configuration	Retrieval hit@10
64–16	0.436
128–32	0.511
256–64	0.580
512–128	**0.615**

Table 3: Retrieval hit@10 for different chunk configurations on PharmaQA.IT (TOPK = 10).

Model	F_1 (ROUGE-L)	EM
Gemini-2.5-Flash	0.329	0.005
Mistral-Large	0.477	0.237
GPT-4.1	0.636	0.367
DeepSeek-Chat	**0.679**	**0.432**
Qwen-Plus	0.579	0.328

Table 4: Closed-source models on PharmaQA.IT with chunking configuration 512–128. We report ROUGE-L F_1 and Exact Match (EM).

nor does it include other languages. Models evaluated on PharmaQA.IT may therefore not generalise to broader biomedical domains or multilingual settings without additional adaptation.

Dataset size and distribution. The final corpus contains 861 question–answer pairs over 1,077 RCPs. This scale is sufficient for robust benchmarking but is relatively small for training or fine-tuning large language models from scratch. Moreover, although we sample across different therapeutic classes, the distribution of topics (e.g., indications vs. pharmacokinetics) and answer types is not perfectly balanced, and some information needs are under-represented. PharmaQA.IT should thus be primarily seen as an evaluation resource rather than as a standalone training set.

Semi-automatic annotation pipeline. Question–answer pairs are generated through a semi-automatic pipeline that relies on a specific multimodal LLM (Qwen3-VL-235B-A22B) to propose candidates, followed by expert validation. This design improves efficiency but introduces potential biases: the style and granularity of the questions may partially reflect the underlying model, and subtle errors could persist despite human checking. In addition, validation was performed by a small pool of expert annotators, which may limit the diversity of perspectives on what constitutes a "natural" or "useful" question.

Task design and evaluation metrics. PharmaQA.IT focuses on extractive QA with short, factoid-style answers grounded in a single RCP. More complex scenarios such as multi-hop reasoning across sections or documents, unanswerable questions, patient-oriented formulations, or generative explanations are not explicitly covered. On the evaluation side, we rely on Exact Match and ROUGE-L F_1, which, while standard, do not fully capture semantic adequacy, calibration, or safety aspects of the answers. We also fix a single retrieval model and configuration (e.g., multilingual-e5-base, top-k chunks), and do not explore alternative retrievers or long-context setups, which may affect absolute performance.

Lack of explicit multimodal supervision. Although the original RCPs are long PDF documents with rich layout, tables, and occasional figures,

PharmaQA.IT is currently framed as a *text-only* extractive QA benchmark. During QA generation, we explicitly instruct the multimodal LLM to ignore figures, plots, and tables, and our evaluation pipeline operates on textual chunks only. As a result, tasks that genuinely require interpreting visual structure (e.g., reading dosage tables or graphical summaries) are not covered, and models are not evaluated on their ability to jointly exploit text, layout, and visual cues. Extending PharmaQA.IT with aligned multimodal annotations—for instance, by linking questions to page images, table regions, or layout-aware spans—is a natural direction for future work towards visual and multimodal QA over regulatory documents.

Acknowledgements

The work was partially funded by:

- Villanova, a project financed by IPICEI-CIS, Prog. n. SA. 102519 - CUP B29J24000850005 [5].
- "ReSpiRA - REplicabilità, SPIegabilità e Ragionamento", a project financed by FAIR, Affiliated to spoke no. 2, falling within the PNRR MUR programme, Mission 4, Component 2, Investment 1.3, D.D. No. 341 of 03/15/2022, Project PE0000013, CUP B43D22000900004 [6];
- "MAESTRO - Mitigare le Allucinazioni dei Large Language Models: ESTRazione di informazioni Ottimizzate" a project funded by Provincia Autonoma di Trento with the Lp 6/99 Art. 5:ricerca e sviluppo, PAT/RFS067-05/06/2024-0428372, CUP: C79J23001170001[7];

References

Daniel Alves and Luisa Coheur. 2022. Bootstrapped distant supervision for named entity recognition. In *Proceedings of LREC*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020a. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637.

Mikel Artetxe and 1 others. 2020b. Xquad: A cross-lingual question answering dataset. In *EMNLP*.

Pierpaolo Basile and 1 others. 2020. Kind: A dataset for italian ner in social media. In *Proceedings of EVALITA*.

Valerio Basile and 1 others. 2012. Evalita 2012 overview. In *Proceedings of EVALITA*.

Valerio Basile and 1 others. 2016. Evalita 2016 overview. In *Proceedings of EVALITA*.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379.

Asma Ben Abacha and 1 others. 2019. Overview of the mediqa 2019 shared task on summarization and inference in medical texts. In *ACL BioNLP Workshop*.

Cristina Bosco. 2000. Towards a treebank of italian. In *Proceedings of LREC*.

Jonathan Clark and 1 others. 2020a. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *TACL*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020b. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *ICML*.

Gautier Izacard and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval-augmented language models. In *ICLR*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020a. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Holger Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

J. Li and 1 others. 2016. Biocreative v cdr task corpus. In *Proceedings of BioCreative*.

Bernardo Magnini and 1 others. 2006. I-cab: the italian content annotation bank. In *Proceedings of CLiC-it*.

[5]Villanova: https://www.opencup.gov.it/portale/web/opencup/home/progetto/-/cup/B29J24000850005

[6]RESPIRA: https://www.opencup.gov.it/portale/web/opencup/home/progetto/-/cup/B43D22000900004

[7]MAESTRO: https://www.opencup.gov.it/portale/web/opencup/home/progetto/-/cup/C79J23001170001

Telmo Menezes and Benjamin Roth. 2019. Distant supervision for ner: A systematic study. *arXiv*.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368.

Chris Quirk and 1 others. 2016. Distant supervision for clinical ner. *Journal of Biomedical Informatics*.

N. Ringland and 1 others. 2019. Nne: A distantly supervised named entity dataset. In *Proceedings of ACL*.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, and 1 others. 2015a. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

George Tsatsaronis and 1 others. 2015b. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*.

Jiawei Zhou, Yu Su, Yijia Wang, Junghyun Chung, Chen Li, Huan Chen, and Xiang Ren. 2023. Universalner: A universal toolkit for cross-domain and multilingual named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Andrea Zugarini and Leonardo Rigutini. 2025a. Pharmaer.it: An italian dataset for entity recognition in the pharmaceutical domain. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 1171–1180. CEUR Workshop Proceedings.

Andrea Zugarini and Leonardo Rigutini. 2025b. Pharmaer.it: an italian dataset for named entity recognition in the pharmaceutical domain. In *Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025).*, Cagliari, Italy. CEUR Workshop Proceedings.

Andrea Zugarini, Kamyar Zeinalipour, Achille Fusco, and Asya Zanollo. 2024a. ECWCA - educational CrossWord clues answering: A CALAMITA challenge. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1239–1244, Pisa, Italy. CEUR Workshop Proceedings.

Andrea Zugarini, Kamyar Zeinalipour, Surya Sai Kadali, Marco Maggini, Marco Gori, and Leonardo Rigutini. 2024b. Clue-instruct: Text-based clue generation for educational crossword puzzles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3347–3356, Torino, Italia. ELRA and ICCL.

A Prompt used for generating QA

Table 5: Prompt for the creation of example question-answer pairs from textual passages, simple variant.

```
QA_GENERATION_FOR_TEXT_SIMPLE_PROMPT = '''Devi generare esempi per un dataset di Visual Question Answering.
Ti viene data in input una pagina di un pdf, convertita come immagine.
Formula una coppia domanda-risposta basandoti sul testo contenuto nell'immagine.
Se ci sono figure, grafici o tabelle dentro alla pagina, ignorale.
Scrivile come se fossi la seguente persona:
{persona}

Note:
1. La domanda deve essere chiara e semplice e riguardare un'informazione ben circoscritta.
2. La risposta deve essere secca.
3. Genera un JSON contenente la domanda (question) e la sua risposta (answer) senza aggiungere altro.
4. Rispetta il seguente schema JSON: {{"question": "", "answer": ""}}.
```

Table 6: Prompt for the creation of example question-answer pairs from **tables**, simple variant.

```
QA_GENERATION_FOR_TABLE_SIMPLE_PROMPT = '''Devi generare esempi per un dataset di Visual Question Answering.
Ti viene data in input una pagina di un pdf, convertita come immagine.
L'immagine contiene una o più tabelle.
Formula una coppia domanda-risposta basandoti sul contenuto di una di queste tabelle.
Scrivile come se fossi la seguente persona:
{persona}

Note:
1. La domanda deve essere chiara e semplice e riguardare un'informazione ben circoscritta.
2. La risposta deve essere secca.
3. Genera un JSON contenente la domanda (question) e la sua risposta (answer) senza aggiungere altro.
4. Rispetta il seguente schema JSON: {{"question": "", "answer": ""}}.
'''
```

B Example of Evaluated QA pairs

Table 7: Example of Evaluated QA pairs.

```
Example of accepted QA pair:
Q: Quali sono gli effetti del ramipril nei pazienti con compromissione renale?
A: Nei pazienti con compromissione renale, l'escrezione renale di ramiprilato è ridotta e le concentrazioni plasmatiche
   di ramiprilato sono elevate, riducendosi più lentamente rispetto ai pazienti con funzione renale normale.

Example discarded because of incorrect information:
Q: Qual è il rischio di tromboembolia venosa (TEV) per 10.000 donne che usano un contraccettivo ormonale
   combinato contenente drospirenone?
A: tra 9 e 12

Example discarded because of incomprehensible formulation:
Q: Qual è la controindicazione per l'uso di Metformina in pazienti con GFR inferiore a 30 ml/min?
A: Metformina è controindicata.
```

C Prompt used for RAG evaluation

Table 8: System prompt used for RAG-based extractive QA evaluation.

```
TEXT_SYSTEM_PROMPT = """
Sei un assistente estrattivo. Rispondi SOLO usando il CONTENUTO nel blocco "Context".
Se l'informazione non è nel contesto, rispondi esattamente: "Non presente nel contesto".
Usa l'italiano. NON inventare nulla. NON fare deduzioni esterne.

DEVI RESTITUIRE SOLO JSON **VALIDO** (senza testo aggiuntivo prima o dopo) con questo schema ESATTO:
{
  "answer": "<risposta concisa>",
  "evidence": [
    {"chunk_id": <numero>, "quote": "<frase esatta dal contesto>"}
  ],
  "chunks_used": [<numeri>],
  "status": "ok|non_present|ambiguous"
}

Regole:
- Cita frasi brevi esatte dal testo come "quote".
- I "chunk_id" si riferiscono alle etichette [Chunk N] nel Context.
- Se trovi più valori in conflitto, usa "status":"ambiguous" e includi tutte le citazioni rilevanti.
- Se non trovi nulla, usa "status":"non_present" e "answer":"Non presente nel contesto".
"""
```