The UNLP 2026 Shared Task on Multi-Domain Document Understanding

Volodymyr Sydorskyi[1,3], Nataliia Romanyshyn[2], Roman Kyslyi[3], Olena Nahorna[4],
[1]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"
[2]Texty.org.ua
[3]Kyiv School of Economics
[3]Preply
v.sydorskyi@kpi.ua, nataliia.romanyshyn@texty.org.ua, rkyslyi@kse.org.ua lena.kobzar@gmail.com

Abstract

This paper presents the results of the UNLP 2026 Shared Task on Multi-Domain Document Understanding. This Shared Task aims to challenge and assess AI capabilities to find the right information in a stack of domain-specific documents and generalize across domains. Participants were required not only to select the correct answer, but also to localize it by predicting the corresponding document and page. A total of 54 teams registered for the competition, 15 teams submitted systems, and 513 runs were evaluated on a hidden test set via Kaggle in a code-only submission format under constrained computational resources. The Kaggle leaderboard is left open for further submissions. Summarizing the contributions of this work, we establish a Ukrainian multi-domain document understanding benchmark, which consists of: (1) a collected dataset; (2) a proposed evaluation metric; and (3) an analysis of top-performing systems evaluated under a unified framework.

1 Introduction

Working with long and complex documents remains a challenging problem in natural language processing (NLP). While recent large language models (LLMs) have made strong progress on question answering tasks (Ma et al., 2025), they still struggle when required to search through document collections and reliably point to the exact source of the answer (Huang et al., 2025). Moreover, comprehensive benchmarks for evaluating document understanding capabilities remain limited for mid-resource languages such as Ukrainian, where underrepresentation in model pretraining and scarce training resources make robust system development challenging.

To address these challenges, the Fifth Ukrainian NLP Conference (UNLP 2026) organized a Shared Task on Multi-Domain Document Understanding[1]. The goal of the task is to evaluate systems that can both retrieve relevant information from domain-specific documents and use it to answer questions. Unlike standard question answering benchmarks, participants were also required to indicate where the answer comes from by predicting the corresponding document and page.

The shared task is formulated as a multiple-choice question answering problem over collections of domain-specific PDF documents. Given a question and six candidate answers, systems are required to:

1. identify the correct answer from six options
2. specify which document contains the answer
3. pinpoint the exact page number where the answer is located

The dataset spans multiple domains with distinct structures and writing styles, including sporting competition rules, medical product instructions, and military regulations. The military domain was kept hidden from participants to test how well systems generalize to previously unseen data.

The competition was hosted on Kaggle[2] in a code-only submission format under fixed computational constraints, encouraging participants to build efficient and reproducible solutions. A total of 54 teams registered, of which 15 actively participated, producing 513 submissions. Participants explored a range of approaches, from relatively simple hybrid retrieval pipelines to more complex multi-stage systems with reranking and instruction-tuned language models.

The main contributions of this shared task are as follows:

- Introduced a dataset for document understanding in Ukrainian, covering three domains, including one hidden.

[1] https://unlp.org.ua/shared-task/

[2] https://www.kaggle.com/t/3ab59dd1807746c99d0a5c3b72580a8b

Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), pages 249–259

- Proposed an evaluation framework for assessing system performance on multi-domain document understanding.
- Provided a comprehensive analysis of top-performing Shared Task systems, evaluated under a unified framework using the proposed dataset and metric.

The remainder of this paper is organized as follows. Section 2 reviews previous work. Section 3 outlines the UNLP 2026 shared task setup, presents the dataset and describes the evaluation metric. Section 4 reports the leaderboard results and summarises the submitted systems. Section 5 concludes the paper, while Section 5 provides an ethics statement, and finally, current limitations are stated.

2 Related Work

Document understanding and question answering (QA) over complex sources have been widely studied in NLP.

Early benchmarks such as SQuAD (Rajpurkar et al., 2016) established reading comprehension as a core NLP task, but focused on short Wikipedia passages. TAT-QA (Zhu et al., 2021) extended this to financial documents combining text and tables, while MultiReQA (Guo et al., 2021) highlighted the difficulty of cross-domain retrieval-based QA. In the visual document understanding space, DocVQA (Mathew et al., 2021a) introduced QA over single-page industry documents, later extended to the multi-page setting by MP-DocVQA (Tito et al., 2023). SlideVQA (Tanaka et al., 2023) expanded multi-page document understanding to slide decks, requiring evidence selection across multiple images alongside answer generation. DUDE (Landeghem et al., 2023) further introduced multi-domain and cross-domain generalization as explicit evaluation goals across visually-rich documents from diverse industries and origins. Our shared task builds on these ideas by combining answer selection with document and page localization but applies them to text-based multiple-choice QA over domain-specific PDF documents in Ukrainian.

Benchmarks for document understanding remain scarce for Ukrainian. Recent efforts have focused on building large language models such as Lapa (Paniv et al., 2025) or MamayLM (Yukhymenko et al., 2025), and Syromiatnikov and Ruvinskaya (2024) explored context-based QA using zero-shot and few-shot LLMs on general-domain texts. The UNLP workshop series has progressively built evaluation infrastructure for Ukrainian: previous shared tasks addressed LLM instruction-tuning (Romanyshyn et al., 2024) and social media manipulation detection (Kyslyi et al., 2025).

The present task extends this line with a more complex scenario: multi-domain document understanding with answer localization over domain-specific PDFs, including a hidden domain to test cross-domain generalization.

Several evaluation frameworks have been proposed for measuring performance in document retrieval and question answering tasks. Mean Reciprocal Rank (MRR) (Voorhees, 1999) and Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2002) are widely used in information retrieval to assess ranking quality, but they do not account for localization within a retrieved document. RAGAS (Es et al., 2024) introduced a suite of reference-free metrics for retrieval-augmented generation, covering faithfulness and answer relevance, yet it does not evaluate page-level precision. In the document understanding space, metrics used in DocVQA (Mathew et al., 2021b) and MP-DocVQA (Tito et al., 2023) focus primarily on answer correctness via Average Normalized Levenshtein Similarity (ANLS), without penalizing incorrect source attribution. Our proposed metric addresses this gap by jointly rewarding answer accuracy, document retrieval correctness, and page-level localization in a single unified score, making it better suited for multi-document, multi-domain settings where source traceability is essential.

3 Task Description

The main goal of the shared task was to build a system capable of retrieving the correct answer to a particular question from multiple-choice options (6 options) based on a set of documents. In addition, the developed system should also provide localization of the corresponding information within the document. Thus, the task can be decomposed into two goals:

- Find the correct answer to a multiple-choice question
- Identify the document and page where the answer was found

Based on the obtained dataset, the system was evaluated across three domains of PDF documents in Ukrainian. Additionally, the third domain was completely hidden from participants during the system development stage, which was intended to ensure robustness of the system to new domains.

3.1 Data

The dataset consists of three types of open-source documents in Ukrainian: sporting competition rules, medical product instructions, and military regulations. In addition to the documents, it includes a set of multiple-choice questions associated with them, along with references to the specific document pages from which the questions were derived (see Appendix B for a data sample example). All documents were provided in PDF format. Each domain was accompanied by a README file describing the content of the documents in both English and Ukrainian (see an example in Appendix A). The documents were sourced from official government resources, such as the Official Portal of the Ministry of Youth and Sports of Ukraine[3], the regulatory documents website of the Ministry of Health of Ukraine[4], and the official web portal of the Parliament of Ukraine[5].

The **sporting rules** are long documents that define competition regulations, including gameplay rules, organizational procedures, participant requirements, officiating, penalties, and integrity measures. Some documents also include ethical provisions and adaptations for people with disabilities. Appendices often contain tables and diagrams.

The **medical instructions** are shorter texts that describe the safe and effective use of medicinal products. They follow a standardized format covering composition, pharmacology, indications, contraindications, dosage, side effects, storage conditions, and regulatory and manufacturer information.

The **military regulations** are extensive documents that define the principles, structure, and functioning of Ukraine's defense sector. They include strategic-level documents outlining defense policy and reform priorities, as well as statutes of the Armed Forces that regulate military service.

3.2 Data Annotation Process

In order to generate questions, we adopted a hybrid approach: a pool of volunteers created questions with multiple-choice answers, and additionally, we generated questions using GPT-4o mini. The latter were validated by volunteers as well as professional Ukrainian linguists to ensure grammatical correctness and eliminate hallucinations. Specifically, we designed an annotation process to assess several aspects: (1) whether the question is accurate, (2) whether the answer options are plausible, and (3) whether the referenced page is correct. For more details, see Appendix C.

Annotators assessed each question along with answers as valid, flagged it for removal, or marked it as uncertain. As 65% of the data was annotated by professional linguists, their judgements dominated the overall data quality profile. Professional linguists applied markedly stricter standards than volunteers, both in terms of rejection and correction. As shown in Table 1, linguists rejected items at nearly twice the rate of volunteers (33.6% vs. 19.4%), a statistically significant difference ($z = 3.24$, $p < 0.01$). Among items approved as valid, linguists further edited at least one field—question phrasing, correct answer, or a page number—in 82.7% of cases, compared to 63.3% for volunteers ($z = 3.77$, $p < 0.001$). These results suggest that professional linguists not only removed a larger share of items but also actively improved the quality of those they retained. The comparison is based on a single domain-matched subset reviewed by both groups ($n = 412$), ensuring that any observed differences in annotation quality are not confounded by domain variation; therefore, the results should be treated as exploratory.

As a result, we obtained a dataset that was partitioned into three subsets: `train` (or `dev`), `public test`, and `private test`. The dataset statistics are presented in Table 2. It is important to note that questions were not written for all documents (as illustrated in the fourth column of Table 2). Additional documents in the training set can be used for system optimization, while additional documents in the test set can be used for more robust system validation.

[3] https://mms.gov.ua
[4] https://mozdocs.kiev.ua/
[5] https://zakon.rada.gov.ua/

Group	Total items	Drop rate	95% CI (drop rate)	Valid items	Correction rate (valid)
Professional linguists	226	33.6%	[27.5%, 39.8%]	150	82.7%
Volunteers	186	19.4%	[13.7%, 25.0%]	147	63.3%

Table 1: Annotation quality metrics by annotator group. Drop rate: proportion of items flagged for removal out of all reviewed items. Correction rate (valid): proportion of approved items with at least one edited field (question, correct answer, or page number).

Domain	Split	#Docs	#Docs w Qs	Avg. len.	#Qs
Sport	train	11	11	78.64	20
	public	19	19	78.21	319
	private	46	46	76.83	704
Medical	train	30	20	8.53	20
	public	50	50	8.26	464
	private	120	106	8.32	942
Military	private	5	5	103	500

Table 2: Dataset statistics.

3.3 Evaluation

$$\begin{aligned}\text{Metric} = 0.5\frac{1}{N}\sum_{i=1}^{N} a_i \\ + 0.25\frac{1}{N}\sum_{i=1}^{N} d_i + 0.25\frac{1}{N}\sum_{i=1}^{N} p_i\end{aligned} \tag{1}$$

$$\begin{aligned}a_i = \mathbb{I}\Big(&\text{Correct\_Answer}_i^{(\text{pred})} \\ &= \text{Correct\_Answer}_i^{(\text{true})}\Big)\end{aligned} \tag{2}$$

$$\begin{aligned}d_i = \mathbb{I}\Big(&\text{DocID}_i^{(\text{pred})} \\ &= \text{DocID}_i^{(\text{true})}\Big)\end{aligned} \tag{3}$$

$$\begin{aligned}p_i = &\left(1 - \frac{\left|\text{Page\_Num}_i^{(\text{pred})} - \text{Page\_Num}_i^{(\text{true})}\right|}{\text{n\_pages}_i}\right) \\ &\cdot \mathbb{I}(d_i = 1)\end{aligned} \tag{4}$$

Final evaluation metric[6] is shown on Equation 1. And it is composed of three terms:

- Answer accuracy - Equation 2
- Document correctness - Equation 3
- Page correctness - Equation 4

The metric is constructed in such a way that its upper bound is equal to one, while the lower bound is theoretically unbounded due to the p_i term. However, by applying a straightforward rule that limits the number of predicted pages for a particular document by the maximum number of pages (n_pages_i), the minimum value is also effectively bounded by zero.

Exploring the first term - Answer accuracy - it is assigned the largest coefficient. This is motivated by the fact that the two subsequent terms reflect the system's localization ability, and we treat them as a single second goal of the task. It is also important to mention that accuracy can be affected by class imbalance; however, since the distribution of correct choices can be controlled, it can always be adjusted to be approximately uniform.

Document correctness is required to evaluate how accurately the system can select the relevant document. For example, this is important in cases where the system needs to retrieve information about a drug that has a very close substitute. It also directly affects the next term of our metric - Page correctness - because selecting any page from an incorrect document is not meaningful.

Finally, Page correctness reflects how close the predicted page is to the correct page, but only if the document has been selected correctly. To measure this closeness, $1 -$ Relative Absolute Error at the page level was used. The motivation for this choice is as follows: using a strict indicator function would be too restrictive, since information relevant to a particular question may be distributed across several pages, or the exact page may not be predicted correctly. In such cases, localization near the correct page is still a more desirable outcome than predictions far from it. Regarding the use of absolute error instead of a quadratic penalty, this choice was made to keep the metric simpler and because it is not obvious how strongly outliers should be penalized.

[6] https://www.kaggle.com/code/vladimirsydor/unlp-2026-document-understanding-metric

Rank	Team	Public Score	Private Score
1	GA	**0.9460**	**0.9598**
2	Bullseye Emoji	0.9211	0.9420
3	Dialogus ex Machina	0.9402	0.9411
4	Golden Retrievers	0.8876	0.9191
5	Daria Belozor	0.8715	0.8802
6	lawandia	0.8528	0.8775
7	PFW	0.8688	0.8722
8	OM	0.8376	0.8314
9	OksanaTkach	0.7720	0.8210
10	catdlia	0.7831	0.8095

Table 3: Leaderboard for UNLP 2026 Shared Task. Final ranking is based on private leaderboard scores; public scores are shown for comparison.

3.4 Data Split and Competition Setup

The Shared Task was conducted on the Kaggle platform[7] using a special type of competition - a Code Competition. This competition format allows all testing data to be hidden from participants, including both target values - correct answer choices, document IDs, and page numbers - and input data - questions, documents, and domain descriptions. This feature is critical for document understanding and information retrieval shared tasks, as it explicitly limits participants from:

- manually finding answers and localizations for test questions,
- overfitting the system to particular documents or questions,
- overfitting the system to all domains available in the test set.

Additionally, all data from the Military domain was placed in the private test set, so participants were not able to explicitly track performance on it via the public leaderboard and ranking. The first two domains were split document-wise in order to avoid data leakage. Detailed data split statistics can be found in Table 2.

4 Results and System Descriptions

In this section, detailed metric results from the top-ranking team, together with a brief description of the top two systems on the public leaderboard, are discussed.

4.1 Overall Results Summary

Metric scores obtained by the top 10 teams on the public and private test sets are presented in Table 3. The overall average absolute deviation between public and private scores is approximately 0.018, with private scores being slightly higher. This deviation is relatively small, indicating consistent performance across the two evaluation settings. Team rankings are also largely preserved, with only minor one-position swaps observed between three pairs of teams.

The results presented in Table 4 reveal several consistent tendencies across domains, metrics, and evaluation settings. First, performance on the private test set is generally higher than on the public test set, indicating that the systems do not exhibit significant overfitting to the public subset and are able to generalize reasonably well to unseen data. This trend is particularly evident for the Sport domain, where improvements are observed across all metrics.

Across domains, the Medical domain demonstrates the highest overall performance, with consistently strong results for all metrics. This suggests that the corresponding documents are either more structured or easier to interpret for retrieval and localization tasks. In contrast, the Sport domain appears more challenging, especially in terms of document and page correctness, indicating potential ambiguity in document selection and localization.

Analyzing the decomposed metrics, answer accuracy is generally high and stable across domains, confirming that systems are effective at selecting the correct answer once relevant information is identified. Document correctness also achieves strong performance, particularly in the Medical and Military domains, suggesting reliable document retrieval. However, page correctness consistently lags behind the other metrics, highlighting the difficulty of precise localization within documents. This gap indicates that, while systems can often identify the correct document, accurately pinpointing the exact page remains a more challenging task.

Finally, the Military domain, available only in the private test set, shows competitive performance levels, which suggests that the systems are robust to domain shifts despite the absence of explicit exposure during development. Overall, the observed trends indicate that the primary bottleneck of current approaches lies in fine-grained localization rather than answer selection or document retrieval.

[7] https://www.kaggle.com

Team	Metric		Answer		Document		Page	
	Pub	Priv	Pub	Priv	Pub	Priv	Pub	Priv
Sport								
GA	0.9195	**0.9481**	0.9279	**0.9574**	**0.9310**	**0.9560**	**0.8914**	**0.9216**
Bullseye Emoji	0.8601	0.9163	0.8840	0.9403	0.8558	0.9091	0.8166	0.8753
Dialogus ex Machina	**0.9198**	0.9220	**0.9310**	0.9176	**0.9310**	0.9503	0.8860	0.9024
Golden Retrievers	0.7675	0.8254	0.8683	0.8991	0.6897	0.7741	0.6435	0.7292
Daria Belozor	0.7981	0.8452	0.8777	0.9048	0.7524	0.8196	0.6845	0.7515
Medical								
GA	0.9629	0.9642	0.9547	**0.9628**	**0.9871**	**0.9841**	0.9551	0.9471
Bullseye Emoji	0.9607	0.9634	0.9569	**0.9628**	0.9763	0.9756	0.9528	0.9523
Dialogus ex Machina	0.9533	0.9528	0.9461	0.9597	0.9828	0.9735	0.9382	0.9186
Golden Retrievers	**0.9703**	**0.9655**	**0.9677**	**0.9628**	**0.9871**	0.9830	**0.9587**	**0.9535**
Daria Belozor	0.9220	0.9075	0.9483	0.9406	0.9440	0.9151	0.8474	0.8337
Military								
GA	–	**0.9684**	–	0.9540	–	**0.9860**	–	**0.9795**
Bullseye Emoji	–	0.9381	–	0.9500	–	0.9300	–	0.9223
Dialogus ex Machina	–	0.9462	–	0.9340	–	0.9700	–	0.9469
Golden Retrievers	–	0.9638	–	**0.9580**	–	0.9760	–	0.9630
Daria Belozor	–	0.8785	–	0.8980	–	0.8820	–	0.8359

Table 4: Performance across 5 top teams, domains, and evaluation metrics. Public (Pub) and private (Priv) scores are reported. Best values per column are highlighted in bold.

4.2 First Place Solution

The winning team created a system that works in three stages: chunks documents, retrieves relevant passages, and then generates an answer (Bazdyrev et al., 2026). First, documents are split into overlapping chunks while keeping important context intact (tables are flattened into text, and each chunk gets a prefix identifying which document it came from). For every multiple-choice option, the system builds an instruction-style query and runs it through dense retrieval using Qwen3-Embedding-8B, pulling the top 20 candidates. A fine-tuned Qwen3-8B reranker then narrows these down to just the top 2 passages. The final answer comes from Qwen3-32B-AWQ, which uses constrained decoding so it can only output a valid option label (A through F). The document and page predictions are simply taken from passage in which reranker scored highest. To train the reranker, the team combined Ukrainian QA data (FIdo-AI/ua-squad) with translated synthetic datasets built from sources like MS MARCO and HotpotQA. Everything was designed to run within Kaggle's hardware limits (two T4 GPUs), relying on vLLM with quantization and trimmed context lengths to fit.

4.3 Second Place Solution

The second place system takes a two-stage approach, pairing hybrid retrieval with a fine-tuned generation model (Nosenko and Kilko, 2026). Documents are first converted to Markdown, then ranked at the document level using both semantic embeddings (computed over the opening text) and BM25. When the two methods disagree, their candidate lists are merged and reranked. Next, the system zooms in to the page level within the chosen document: it chunks pages, scores them with embeddings and BM25, blends the results using Reciprocal Rank Fusion, and applies a custom reranker on top. To boost performance, the team generated roughly 7,000 synthetic QA examples from the training data and used them to fine-tune MamayLM with LoRA. The model learns to predict both the answer option and the page number in one shot (outputting something like "A 2"). For deployment, the model is quantized and packaged in GGUF format, running through llama.cpp. At inference time, it gets the top 3 retrieved pages as context to work with.

5 Conclusion

We believe that the UNLP 2026 Shared Task on Multi-Domain Document Understanding is instrumental in facilitating research on retrieval-augmented question answering for Ukrainian language documents. The task introduced a novel evaluation scenario requiring systems to simultaneously select the correct answer from six options, identify the source document, and pinpoint the exact page, evaluated under a unified metric that jointly rewards answer accuracy and localization quality. The use of a hidden military domain further tested cross-domain generalization under realistic constraints, revealing the ability of top systems to

adapt to previously unseen document types.

A total of 54 teams registered, 15 submitted systems, and 513 runs were evaluated on Kaggle under fixed computational constraints. Teams explored a variety of techniques, from hybrid retrieval combining dense embeddings and BM25 to cross-encoder reranking, synthetic data generation for fine-tuning, and post-answer page verification and demonstrated the creative potential of the NLP research community when working in low-resource settings. Top-performing systems employed models such as Qwen3-32B, Lapa, and MamayLM alongside multilingual retrievers like BGE-M3. All final solutions were required to be released under an open license, promoting reproducibility and accessibility. The Kaggle leaderboard remains open for further submissions.

We hope this shared task will serve as a foundation for future work in Ukrainian NLP, and that the tools, data, and approaches developed through this competition will continue to support progress in document understanding and information retrieval for low-resource languages.

Ethics Statement

To ensure fair competition and support the development of reproducible and transparent solutions, the shared task was conducted under a set of clearly defined rules governing data use, system design, and participant behavior.

By participating in the shared task, all teams agreed to comply with the competition rules. In particular, participants were required to avoid any form of unfair advantage, including leaderboard probing or attempts to infer hidden test data. The hidden domain used for evaluation was explicitly protected, and any effort to retrieve or approximate its content through indirect means was strictly prohibited.

Participants were required to use open-source models in their final solutions, proprietary models were permitted only for data generation purposes. The use of external data was allowed provided the corresponding licenses permitted research use. We encouraged participants to use Ukrainian-specific language models such as Lapa and MamayLM. Final solutions were required to be released under an open license, promoting reproducibility and accessibility for the research community.

The dataset used in the shared task consists of publicly available documents, and no personal or sensitive data was included. Questions were generated and validated through a combination of automated methods and human review, with additional validation to ensure correctness and minimize potential errors.

Limitations

We identify the following limitations of our work:

- The dataset covers only three domains, which consist of relatively well-structured and typical documents. To make the benchmark more challenging and representative of real-world scenarios, it should be extended to a broader range of domains with less structured data, such as newspapers, blogs, and social media posts.
- The proposed evaluation metric was not compared with existing metrics for document retrieval in terms of correlation with subjective or human-centered evaluation criteria. As a result, it lacks sufficient empirical justification. A more thorough comparative analysis would strengthen the validity of the proposed metric.
- The collected benchmarks are primarily aimed at maximizing the proposed metric rather than encouraging a broader diversity of methodological approaches. Consequently, many solutions rely on similar techniques. Expanding the benchmark to promote methodological diversity, rather than optimization of a single metric, would further strengthen the study.

Acknowledgments

We are grateful to the volunteers who assisted in question generation and validation and to Preply that sponsored the validation of the dataset by professional Ukrainian linguists. We also thank the Kaggle platform for hosting the competition infrastructure.

AI-assisted tools (ChatGPT and Claude) were used exclusively to improve the clarity and grammar of this text; they did not contribute to the research design, experiments, or analysis.

References

Anton Bazdyrev, Ivan Bashtovyi, Ivan Havlytskyi, Oleksandr Kharytonov, and Artur Khodakovskyi. 2026. Qwen goes brrr: Off-the-shelf rag for ukrainian multi-domain document understanding. In *Proceedings of*

the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026), Lviv, Ukraine. Association for Computational Linguistics. To appear.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158.

Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2021. MultiReQA: A cross-domain evaluation forRetrieval question answering models. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 94–104, Kyiv, Ukraine. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Roman Kyslyi, Nataliia Romanyshyn, and Volodymyr Sydorskyi. 2025. The UNLP 2025 shared task on detecting social media manipulation. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 105–111, Vienna, Austria (online). Association for Computational Linguistics.

Jordy Van Landeghem, Rafał Powalski, Rubèn Tito, Dawid Jurkiewicz, Matthew Blaschko, Łukasz Borchmann, Mickaël Coustaty, Sien Moens, Michał Pietruszka, Bertrand Ackaert, Tomasz Stanisławek, Paweł Józiak, and Ernest Valveny. 2023. Document understanding dataset and evaluation (dude). In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19471–19483.

Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. 2025. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24578–24597, Suzhou, China. Association for Computational Linguistics.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021a. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021b. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Mykola Nosenko and Pavlo Kilko. 2026. Rag pipeline strategies for ukrainian multi-domain document understanding task. In *Proceedings of the Fifth Ukrainian Natural Language Processing Conference (UNLP 2026)*, Lviv, Ukraine. Association for Computational Linguistics. To appear.

Yurii Paniv, Bohdan Didenko, Mykola Haltiuk, Vladyslav Humennyy, Andrian Kravchenko, Roman Kyslyi, Viktoriia Makovska, Artem Orlovskyi, Bohdan Ruban, Maksym-Yurii Rudko, Anastasiia Senyk, Nazarii Drushchak, Dmytro Chaplynskyi, and Mariana Romanyshyn. 2025. Lapa LLM v0.1.2 — the most efficient Ukrainian open-source language model.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Mariana Romanyshyn, Oleksiy Syvokon, and Roman Kyslyi. 2024. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. In *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 67–74, Torino, Italia. ELRA and ICCL.

Mykyta V. Syromiatnikov and Victoria Ruvinskaya. 2024. Ua-llm: Advancing context-based question answering in ukrainian through large language models. *Radio Electronics, Computer Science, Control*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: a dataset for document visual question answering on multiple images. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Ellen M Voorhees. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Hanna Yukhymenko, Anton Alexandrov, and Martin Vechev. 2025. Mamaylm v1.0: An efficient state-of-the-art multimodal ukrainian llm.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Domain Description Example

This appendix provides an example of a domain description (README file) accompanying the dataset.

This folder contains the rules of sporting competitions for various sports. All documents are written in Ukrainian. The average length of a document is 77 pages.

Each document provides a clear and structured presentation of the rules governing the conduct of competitions, as well as ensuring fairness and the safety of participants.

The main sections of a competition rules document include the following information:

- general points describing the scope of application of the rules;
- terms and abbreviations used in the rules;
- rules of games and competitions, often taking into account different disciplines and formats of the sport concerned;
- competition organization processes and competition regulations;
- requirements for competition venues, equipment, and facilities;
- requirements for competition participants and qualification criteria;
- rules for team formation in team competitions;
- medical supervision of competition participants;
- officiating, scoring, and determination of winners;
- violations of competition rules, penalties, and disqualification;
- participants' protests and appeals against officials' decisions;
- anti-doping rules and measures;
- measures to prevent corruption in competitions.

The rules of some sports also include ethical provisions and adaptations for war veterans, people with disabilities, and individuals with conditions that limit their daily activities.

The documents may include diagrams and tables, most of which are placed in appendices.

The rules of sporting competitions serve as a primary source of information for organizers, officials, and participants, ensuring unified standards and a shared understanding of the rules of play.

B Data Format Example

This appendix illustrates the structure of a single data row from the dataset, clarifying which fields are available to participating systems as input, which must be predicted, and which are technical (additional) ones.

Field Roles

- **Input fields** (visible to the system): `Question_ID`, `Question`, `A`, `B`, `C`, `D`, `E`, `F`
- **Target fields** (output of the system): `Correct_Answer`, `Doc_ID`, `Page_Num`
- **Additional fields**: `Domain`, `n_pages`

C Question Validation Guidelines

This appendix describes the validation procedure for generated questions used in the *UNLP 2026 Shared Task on Multi-Domain Document Understanding*.

C.1 Task Overview

Participants are provided with sets of PDF documents, where each set corresponds to a particular domain. In addition to the PDFs, each set includes two text files, in Ukrainian and English, describing the domain.

For each domain, a set of multiple-choice questions is prepared in a standard quiz format, consisting of one question and six answer options. Participants are required to build a system that can: (i) identify the correct answer, (ii) indicate the exact document and page where the answer was found, and (iii) perform well on these tasks regardless of the domain or document length.

The domains are as follows:

- **domain_1**: rules of sporting competitions across various sports approved by the Ministry of Youth and Sports of Ukraine;
- **domain_2**: medical product instructions;
- **domain_3**: currently undisclosed.

Field	Value (Ukrainian)	Value (English)
Input fields		
Question_ID	0	0
Question	Що означає термін «у грі» на змаганнях з регбі?	What does the term “in play” mean at rugby competitions?
A	гравця позначили маленькою позначкою ’X’	the player was marked with a small ‘X’ marker
B	гравець перебуває в положенні, що дає можливість брати участь у грі	the player is in a position that allows them to participate in the game
C	гравець перетнув лінію вкидання	the player has crossed the throw-in line
D	гравець покараний за положення «поза грою»	the player was penalized for being in an offside position
E	удар ногою, призначений на користь невинної в порушенні команди	a kick awarded in favor of the team not at fault for the infringement
F	гравець опинився ближче 10-ти метрів від суперника	the player ended up within 10 meters of an opponent
Additional fields		
Domain	domain_1	domain_1
n_pages	79	79
Target fields		
Correct_Answer	B	B
Doc_ID	759dfc8486c0f02391d7 cfc1fed753b0608fc601.pdf	759dfc8486c0f02391d7 cfc1fed753b0608fc601.pdf
Page_Num	43	43

Table 5: Example data row split by field role.

C.2 Validation Task

Validators are asked to review each generated question and mark the corresponding value in the Validated? field. The following criteria should be taken into account.

C.2.1 Question Formulation

- The wording of the question should allow one to identify unambiguously the document that contains the answer. For example, it is too broad to ask about clothing in karate in general, since multiple types of karate are represented in the collection. Instead, the question should specify a particular discipline, such as kyokushin karate competitions.

- The question must be answerable using only the content of the given document. No external resources, including internet sources, should be needed.

- Overly trivial questions should be removed. For instance, a question such as “What active substance is contained in caffeine sodium benzoate?” is not informative if the answer simply repeats the name of the product.

- If the name of the medication or sport is omitted from the question, it should be added whenever necessary to avoid ambiguity.

C.2.2 Answer Options

- The answer to each question must be located on a single page of the document. Accordingly, the Page_num field should contain exactly one page number.

- If the relevant information is repeated on multiple pages, such questions should preferably be excluded.

- If the relevant context spans the boundary between two pages, it is better not to formulate a question based on that passage.

- Each question must have exactly one correct answer and five incorrect answers.

- Incorrect answers should be plausible. They may be constructed on the basis of the same document. For example, in a question about active ingredients in a medication, excipients listed in the same instruction may serve as distractors.

- One of the incorrect answers may be intentionally nonsensical in order to test models for hallucinations.
- GPT-generated questions and answers may occasionally contain phrasing errors, lexical mistakes, or other inaccuracies. Such cases should be corrected during validation.

C.2.3 Page Numbering

- In the `Page_num` field, make sure the page number corresponds to the one in the PDF reader rather than the page number printed inside the document itself, if such numbering is present.
- In longer documents, GPT may assign incorrect page numbers. Please correct these errors whenever they are detected.

C.3 Practical Note

For convenience, the generated questions are grouped by document and page. This allows one to work with one document at a time and review all associated questions efficiently.