

# **An End-to-End Ukrainian RAG for Local Deployment. Optimized Hybrid Search and Lightweight Generation**

Mykola Trokhymovych

Yana Oliinyk

Nazarii Nyzhnyk

<https://arxiv.org/pdf/2604.22095>

# Shared Task: Goal & Constraints

**The Task:** Develop a system to answer a question with 6 multiple-choice options and output:

Correct\_Answer, Doc\_ID, and Page\_Num.

**Hardware/Time:** Must run on Kaggle P100 or 2xT4 GPUs within **9 hours** total.

**Environment:** Code-only submission; no internet access allowed during inference.

**Scoring function:**

$$\text{Metric} = 0.5 \cdot \frac{1}{N} \sum_{i=1}^N a_i + 0.25 \cdot \frac{1}{N} \sum_{i=1}^N d_i + 0.25 \cdot \frac{1}{N} \sum_{i=1}^N p_i$$

**where**  $a$  - answer correct (1/0),  $d$  - doc correct(1/0),  $p$  - page proximity (how far is the page from expected)

**Challenge:** Find the answer AND the exact page number.

# Data

- 3 domains:
  - medicines instructions
  - sports official rules and documentation
  - hidden
- input docs: ~240 PDF documents with varying layouts and lengths in Ukrainian.
- questions dataset : questions with choices (A/B/C/D/E/F)
- expected output : letter + doc\_id + page number

ЗАТВЕРДЖУЮ

Заступник Міністра  
молоді та спорту України

Сергій ТИМОФЄЄВ

ПРАВИЛА  
спортивних змагань з регбі

ІНСТРУКЦІЯ  
для медичного застосування препарату

РЕТАБОЛІЛ  
(RETABOLIL)

Склад:  
діюча речовина: надролон;  
1 мл розчину містить надролону деканоату 50 мг;  
допоміжні речовини: спирт бензиловий, спирт ізопропіловий, олія соняшникова.

Лікарська форма. Розчин для ін'єкцій.

Фармакотерапевтична група. Анаболічні засоби для системного застосування. Похідні естрену. Код АТХ A14A B01.

Клінічні характеристики.

Показання.

Період реконвалесценції. Анемія внаслідок хронічної ниркової недостатності.

Протипоказання. Вагітність; період годування груддю; карцинома простати, карцинома грудної залози у чоловіків; нефротичний синдром при тривалій анаболічній терапії; тяжка печінкова недостатність; метастази у пацієнтів з онкозахворюваннями; порфірія; підвищена чутливість (алергія) до будь-якого із компонентів препарату; дитячий вік.

Спосіб застосування та дози.

Анемія.

Дорослим: внутрішньом'язово, чоловікам – по 200 мг 1 раз на тиждень, жінкам – по 100 мг 1 раз на тиждень. Терапевтичний ефект залежить від індивідуальної чутливості пацієнта, характеру захворювання. Лікування варто припинити у разі відсутності терапевтичного ефекту після 3-6 місяців застосування.

Ілля) визначають  
нь з регбі, що  
ил та Положень  
документом для  
рії України.  
авцям, тренерам,  
уть створювати  
ливність гравцям

варто  
них рівня  
вання

якляс

пуклини

Болічних  
воюю  
к ефіри  
чити.

ісіння за  
о циклу,

ій і  
клітора.

ка,

к та як

| Question_ID | Domain   | # n_pages | Question                                                                                                | A                                        | B                                                                   | C                               | D                                          | E                                                               | F                                                  | Correct_A... | Doc_ID                             |     |
|-------------|----------|-----------|---------------------------------------------------------------------------------------------------------|------------------------------------------|---------------------------------------------------------------------|---------------------------------|--------------------------------------------|-----------------------------------------------------------------|----------------------------------------------------|--------------|------------------------------------|-----|
| 0           | domain_1 | 79        | Що означає термін «у гри» на змаганнях з регбі?                                                         | гравця позначили маленькою позначкою 'X' | гравець перебуває в положенні, що дає можливість брати участь у гри | гравець перетнув лінію вкидання | гравець покараний за положення «поза грою» | удар ногою, призначений на користь невинної в порушенні команди | гравець опинився ближче 10-ти метрів від суперника | B            | 759dfc8486c391d7cfc1feb6688fc601.  |     |
| 1           | domain_1 | 59        | Чий голос є вирішальним щодо переможця змагань з веслування на байдарках і каное за відсутності фото... | виконавчого директора змагань            | головного секретаря змагань                                         | учасників змагань               | старшого судді                             | лікаря змагань                                                  | судді-інспектора                                   | D            | cadeff3fc759353b7c94387926b5d92d3. |     |
| 2           | domain_1 | 173       | З якої причини учасник змагань                                                                          | неявка на являвання                      | отримання ушкодження чи                                             | грубе порушення правил          | непроходження тестів та                    | своєчасна явка на змагання                                      | одяг учасника не відповідає                        | E            | 54a634188fe50d79a2653231c8a2       | 134 |

# Mission: Document Understanding



## The Question

Read a query in Ukrainian  
and understand the  
nuance.



## The Document

Identify which specific file  
contains the answer.



## The Page

Pinpoint the exact page for  
evidence grounding.

The question is ...

**Why should this be interesting at all?**

# Challenges

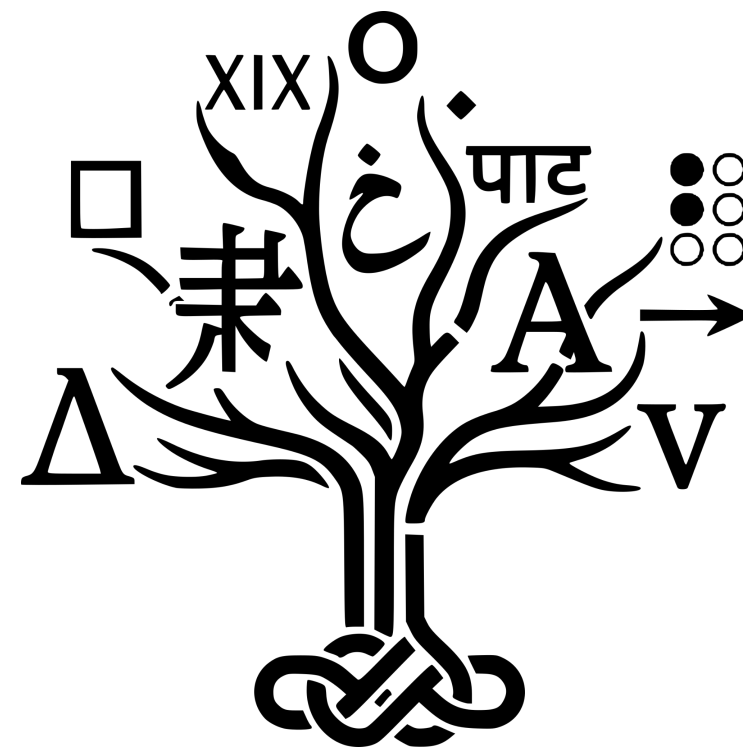
Offline Hunger Games



The Forbidden Fruit



The Linguistic Challenge



# Challenges

## Offline Hunger Games



## The Forbidden Fruit

## The Linguistic Challenge

The main job was not to implement RAG; it was to build it in a **prison cell** (virtually):

-  **No Internet:** No calling GPT-5.2+ for help!
-  **9 Hour Limit:** Process the whole test set in one work day.
-  **Kaggle Hardware:** Only a single P100 or two T4 GPUs.

# Challenges

Offline Hunger Games

The Forbidden Fruit

The Linguistic Challenge

We had training data for Domain 1 & 2. But the **Test Set** included a secret **Domain 3**

If we hard-coded for the training data, we would fail the secret domain. We had to build a system that **actually generalizes**

*Generalization is the name of the game!*



# Challenges

Offline Hunger Games

The Hidden Fruit

The Linguistic Challenge



# RAG: Our Secret Weapon



## Hybrid Search

Combining BM25  
(keywords) with contextual  
embeddings.



## Synthetic QA

Fine-tuning the model on  
synthetic questions  
generated with large LLM!



## Lightweight LLM

Quantizing + fine-tuning  
the model so it fits on  
small GPUs.

# Hybrid Search – The Best of Both Worlds

## Stage 1: Document Hunt

**Goal:** Narrow the candidate pool to the single most relevant document.

**Doc representation:** its first 300 characters

**Dense Scoring:** Uses **pplx-embed-context-v1-0.6b** (0.6B params) for semantic similarity.

**Sparse Scoring:** Employs **BM25Okapi** to index and rank the full preprocessed text.

**Refinement:** If top results disagree, **jina-reranker-v3** evaluates the top two from each method using the first 300 characters + the best BM25 snippet.



## Stage 2: Page Search

**Goal:** Pinpoint the exact chunk and page where the answer resides.

**Syntactic Chunking:** Documents are split into Markdown-based chunks (≤500 chars, 10% overlap) restricted to a single page.

**Hybrid Ranking:** Combines **Cosine Similarity** (via Perplexity embeddings) and **BM25 scores** using

**Reciprocal Rank Fusion (RRF).**

**Final Rerank:** A cross-encoder based on **bge-reranker-v2-m3** reranks the top 8 candidates for final generation.

# Synthetic Dataset Generation - get FT data

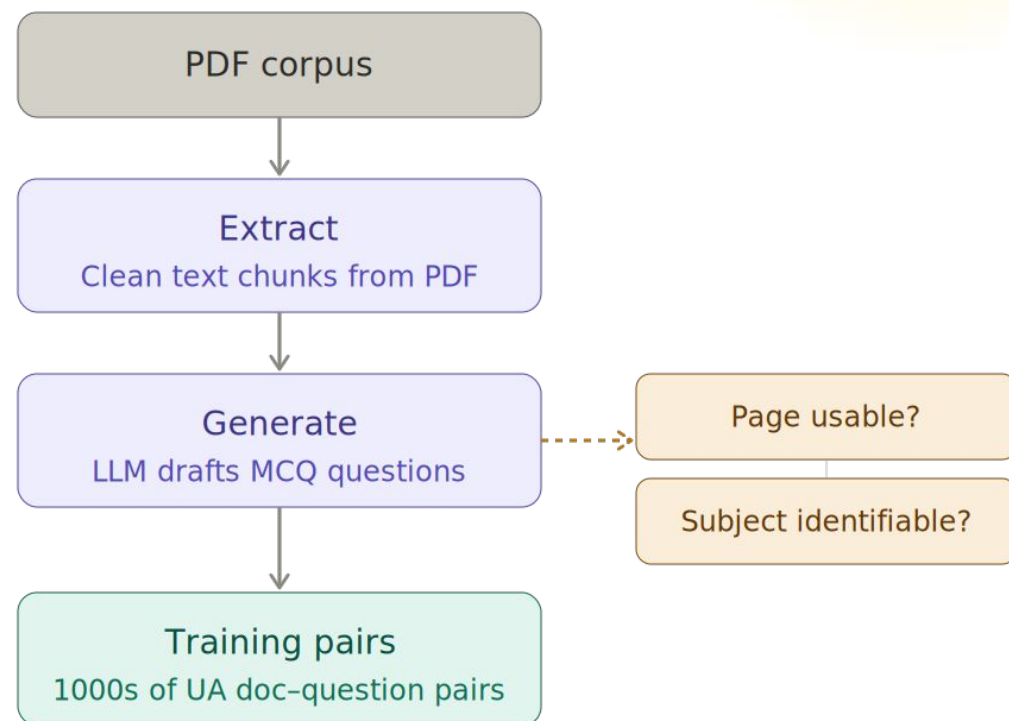
**Cold Start Problem:** With limited initial training pairs, we used a gpt-4o-mini model to generate high-quality synthetic data.

## The Pipeline:

1. **Extract:** Pull clean text chunks from the provided PDF corpus.
2. **Generate:** Use a larger LLM to draft complex multiple-choice questions based on those chunks.

Include validations like: check if page can be used for answering, if the subject is identifiable

**Outcome:** Created thousands of Ukrainian document-question pairs to train our reranker and fine-tune the LLM.



gpt-4o-mini · cold start solved

# Question Generation Prompt

SYSTEM:

You are an expert Ukrainian-language exam question writer.

DOMAIN CONTEXT:

{domain\_description}

YOUR TASK:

You will receive the text of a single page from a Ukrainian PDF document. You must:

1. IDENTIFY the specific subject: the exact sport name (e.g. "strongman", "sambo") or drug name (e.g. "retabolil", "fervex"). This is the ENTITY NAME.
2. DECIDE whether this page contains specific factual content suitable for question generation (specific rules, dosages, penalties, contraindications, etc.).  
SKIP pages that are tables of contents, title pages, abbreviation lists, or general introductions.

3. If suitable, generate up to 10 MCQs that:
  - Are written entirely in Ukrainian.
  - CRITICAL: Every question MUST explicitly name the ENTITY NAME. Generic questions are FORBIDDEN.
  - Are answerable ONLY from the provided page text.
  - Have exactly 6 options (A-F), one correct answer, and 5 plausible distractors.
  - NEVER use quotation marks or braces.
  - Match the style of: {few\_shot\_examples}
4. If NOT suitable, return an empty questions list.

RESPONSE FORMAT (strict JSON):

```
{
  "entity_name": "<sport or drug name>",
  "questions": [
    {
      "question": "...",
      "A": "...", "B": "...", "C": "...",
      "D": "...", "E": "...", "F": "...",
      "correct_answer": "A"
    }
  ]
}
```

Return ONLY valid JSON. "correct\_answer" must be one of: A, B, C, D, E, F.

If not suitable: {"entity\_name": "", "questions": []}

USER:

Domain: {domain}

Document page text: {page\_text}

# LLM selection: Meet the Local Heroes



Lapa LLM 12B



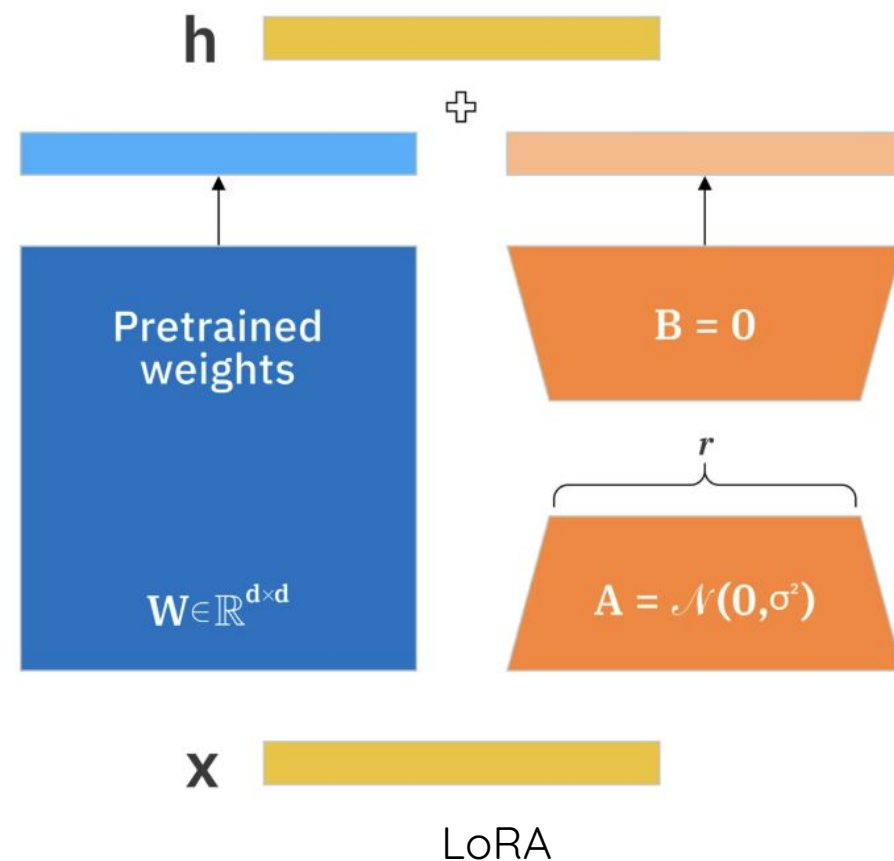
# Optimization – Quantization & Fine-Tuning

**Fine-Tuning (LoRA):** We performed Parameter-Efficient Fine-Tuning (PEFT) on **MamayLLM**. This allowed us to adapt the model to the specific "Ukrainian Document" domain without retraining all 12b+ parameters.

**Quantization (4-bit/AWQ) with llama-cpp:** To fit the model and the retrieval overhead within the Kaggle T4 limits, we quantized the weights to 4-bit precision.

- **Result:** Reduced memory footprint by ~**70%** with minimal loss in reasoning accuracy.

**The Win:** This allowed us to run the entire pipeline (Retrieval + Reasoning) for 240+ documents well within the 9-hour limit.



# Answer Generation Prompt

## Given:

- context chunks
- question

## Generate:

- letter of a correct answer
- page number

## Example of generations:

- "A 1"
- "D 14"
- "E 3"

Context (excerpts from PDF files - each excerpt is separated by ``` characters and contains a page number enclosed in []):

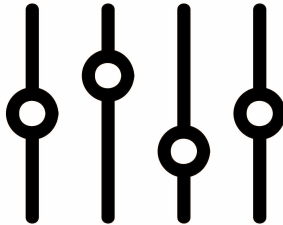
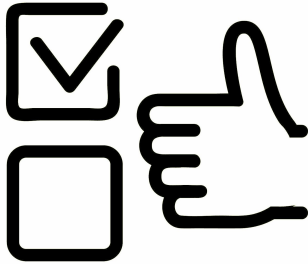
```
```  
Page: [<page_number_1>]  
<page_text_1>  
```  
Page: [<page_number_2>]  
<page_text_2>  
```  
Page: [<page_number_3>]  
<page_text_3>  
```
```

```
Question: <question>  
Options:  
A: <option_A>  
B: <option_B>  
C: <option_C>  
D: <option_D>  
E: <option_E>  
F: <option_F>
```

Instructions:

- Answer the Question using the Context.
- Return the letter of the correct answer (A B C D E F) and the page number where the information was found, separated by a space (e.g., A 1).
- Think carefully; first eliminate the obviously irrelevant options.

# Architecture: End-to-End Pipeline

| Preparation                                                                       | Retrieval                                                                          | Optimization                                                                        | Generation                                                                          |
|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
|  |  |  |  |
| Data ETL                                                                          | Hybrid Search                                                                      | LLM Tuning                                                                          | Verified Answer                                                                     |
| Convert PDFs to Markdown chunks (max 500 chars) with page metadata.               | Two-stage dense/sparse retrieval with RRF and custom BGE reranking.                | Synthetic data generation for 4-bit LoRA fine-tuning.                               | LLM call with context to output verified answers and citations.                     |

# We Came, We Saw, We Ranked #2



 **2nd Place**

Our score actually *improved* on the hidden private data, proving our model didn't just memorize the training set.



**Read the work and Connect on LinkedIn!**



paper



Mykola  
Trokhymovych



Yana  
Oliinyk



Nazarii  
Nyzhnyk

<https://arxiv.org/pdf/2604.22095>