Ministry of Education and Science of Ukraine

NATIONAL UNIVERSITY OF KYIV-MOHYLA ACADEMY

Department of Informatics, Faculty of Informatics

Retrieval Augmented Generation for Ukrainian Government Services: A Comparative Evaluation of RAG approaches

Marynych Anton

Thesis supervisor: Kurochkin Andrew

2025

Context

LLM Usage Stats

- 67% of organizations use generative AI powered by LLMs for content creation.
- 88% of users report improved work quality with LLMs.
- 62% of educational test scores improved with LLM-based quizzes.
- 60% of Bank of America clients rely on LLMs for financial advice.

Source: https://keywordseverywhere.com/blog/llm-usage-stats/?utm_source=chatgpt.com

LLM issues

- Hallucination

- Out of date information

- Cost of fine-tuning

What is RAG?

Retrieval Augmented Generation (RAG) - is an approach for improving the output of the LLM by using the knowledge from external data sources, mainly texts.

Objectives

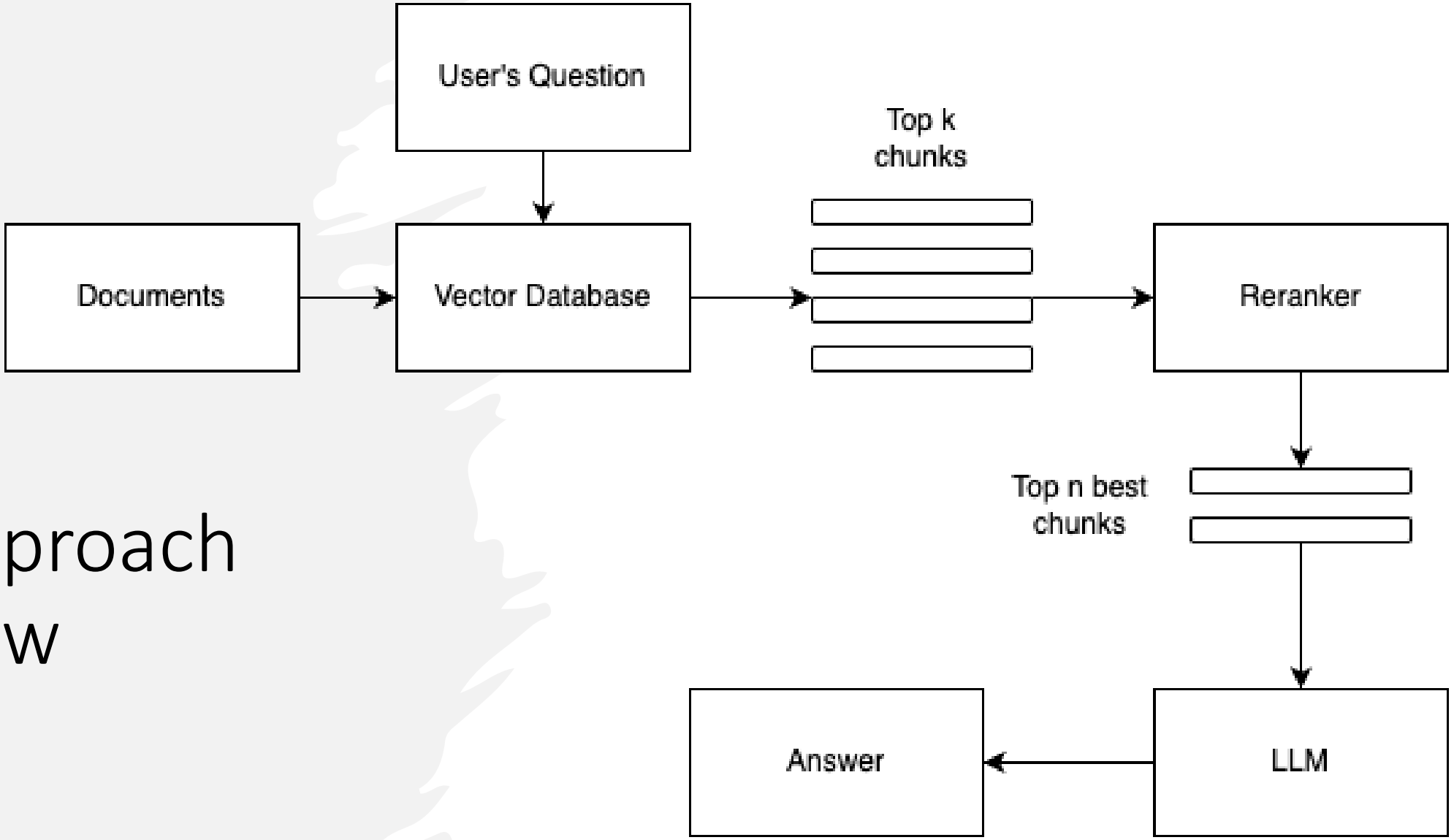
Compare retrieval strategies and architectures for answering questions on Ukrainian government services.

Measure how RAG improves answer accuracy and relevance versus standard LLMs.

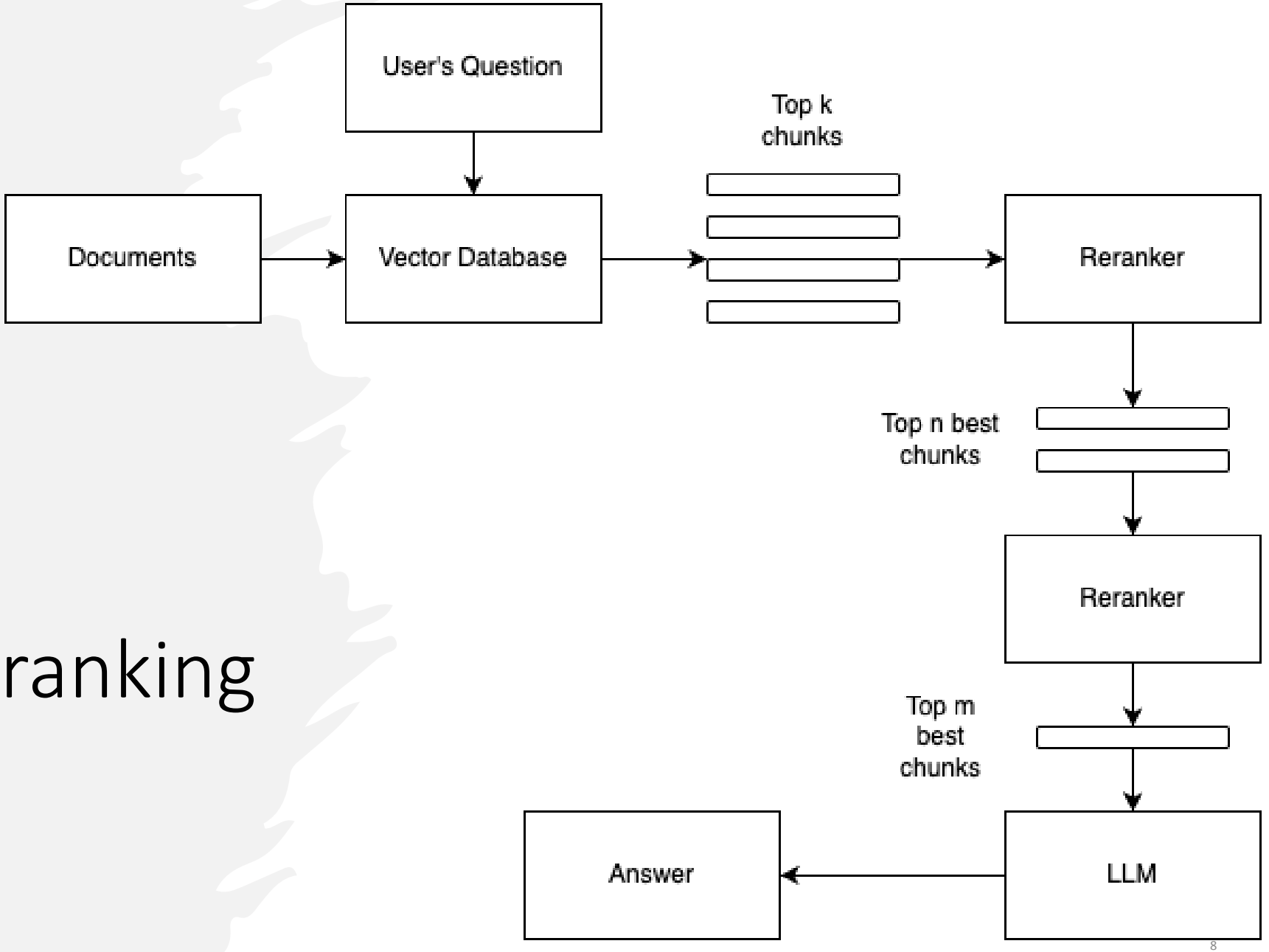
Establish best practices for building and tuning RAG systems in low-resource languages.

Approach

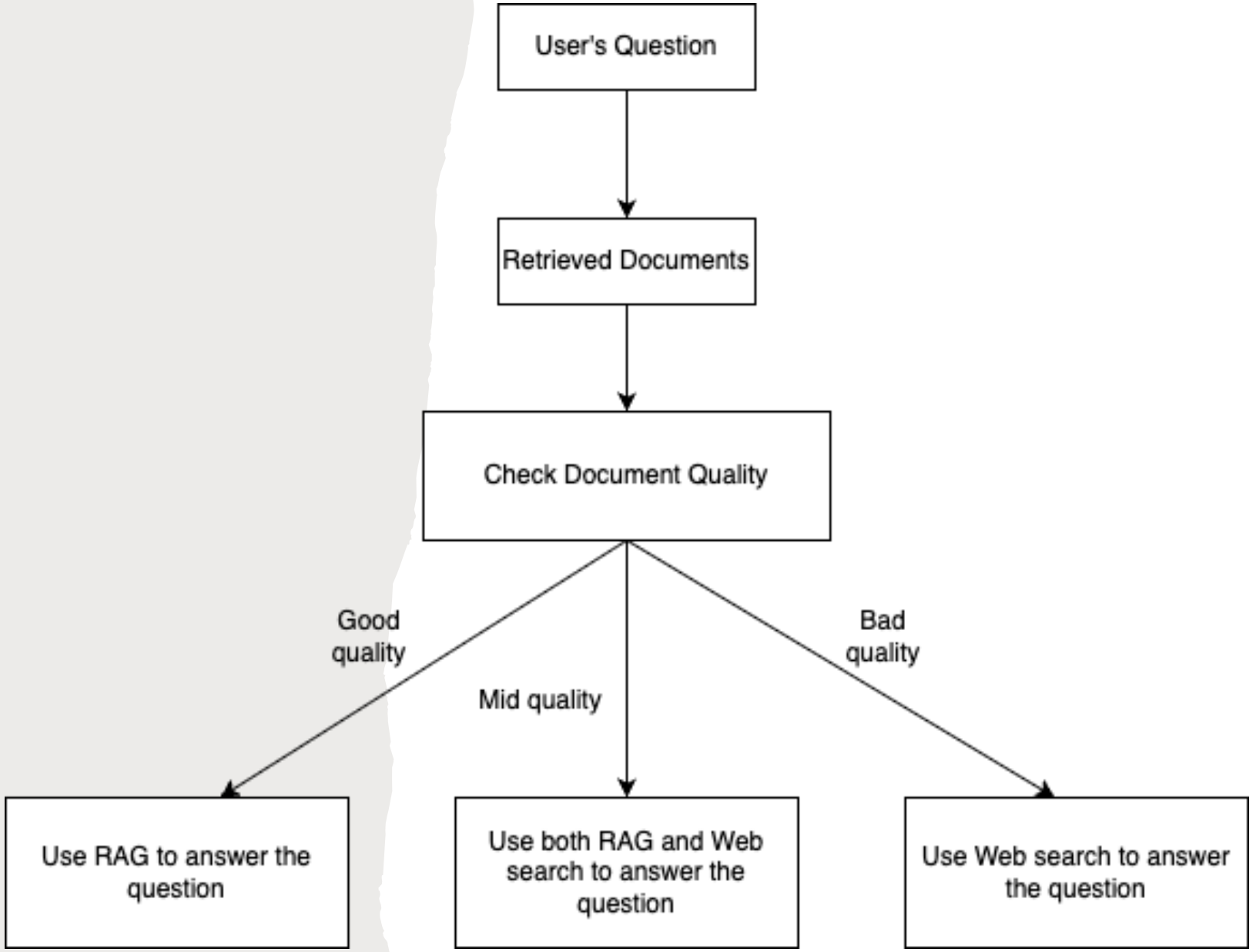
RAG approach overview



RAG with reranking



CRAG


User's Question
Retrieved Documents
Check Document Quality
Good quality
Mid quality
Bad quality
Use RAG to answer the question
Use both RAG and Web search to answer the question
Use Web search to answer the question

Dataset

Documents

- I used the documents from this link: https://guide.diia.gov.ua
- This JSON document has descriptions of 2161 Ukrainian Government Services

Document Description

Some of the properties that each service has:

- Title
- Refusal grounds
- Required documents
- Costs
- Description
- Processing durations
- Receiving ways
- Application ways

Document Snippet


```
"input": [
  "Заява-анкета",
  "Свідоцтво про народження",
  "Оригінали документів, що підтверджують громадянство та особу батьків",
  "Паспорт громадянина України для виїзду за кордон (для повернення з-за кордону)",
  "Посвідчення про взяття на облік бездомних осіб (для бездомних)",
  "Довідка про внутрішньо переміщену особу",
  "Документи для внесення додаткових даних до безконтактного електронного носія (за наявності)",
  "Письмова заява (для написання латиницею прізвища/імені)",
  "Документи для написання латинською (за наявності)",
  "Фотокартка 10×15 см (для осіб з обмеженою мобільністю)"
],
"application_ways": [
  "Заявник: Письмово; Особисто",
  "Представник: Письмово; Особисто"
],
"receiving_ways": [
  "Заявник: Особисто; Письмово",
  "Представник: Особисто; Письмово"
],
"processing_durations": "20 днів (робочі)",
"costs": "Безоплатно",
```

Evaluation Dataset

- JSON file with 500 questions that test the knowledge of the RAG system about Ukrainian Government Services
- The dataset was created using GPT-o4-mini-high model with human evaluation of the quality of the questions
- Each question has an answer written in the dataset

```
{
  "question": "Скільки часу триває оформлення паспорта громадянина України з безконтактним
  електронним носієм уперше після досягнення 14 років?",
  "answer": "Строк звичайного надання становить 20 робочих днів."
},
```

Evaluation approach

Evaluation metrics

- Answer Relevancy
- Semantic Similarity
- Answer Correctness
- Bleau Score
- ROUGE Score
- **Nv Accuracy (LLM judge evaluation)**
- **Factual Correctness**

Main evaluation metrics

Nv Accuracy:
A 0–1 score showing how well a model's answer matches the correct answer by having the model "judge" its own response twice (swapping roles) and averaging the results.

Factual Correctness:
A 0–1 score that splits both the model's answer and the reference into individual factual claims, then measures their overlap (via F1) to see how accurately the response matches the true facts.

Solution and evaluation

Evaluation

Evaluation of the system’s performance was made by inputting all the questions from dataset to the system and calculating the accuracy of system’s outputs.

The evaluation of correctness of system’s answers was done using nv_accuracy and factual_correctness metrics.

Overall Results

Approach	Factual Correctness	LLM Judge Accuracy
GPT-4.1-mini (baseline)	**0.26**	**0.29**
Naïve RAG	0.2675	0.34
Naïve RAG with Reranker	0.32	0.38
SelfRAG with Reranker	0.3424	0.4125
CRAG with Reranker	**0.3731**	0.43
CRAG with Reranker and HyDE	0.3398	**0.4325**

Comparing to existing studies

- **Boros et al. (UNLP 2024):** +10 pp Ukrainian QA accuracy (30.9 → 40.2 %) on ZNO dataset
- **Chirkova et al. (mRAG 2024):** ~+14 pp non-English QA (e.g., Arabic 26.4 → 45.9; Japanese 31.7 → 42.7)
- **This Work:** +11 pp Factual Correctness (F1) • +14 pp LLM Judge (Answer Accuracy)

Conclusions

Conclusions

- This is a first work which used RAG with Ukrainian documents and evaluated it using open answer questions
- We achieved 14% improvement in LLM judge accuracy and 11% improvement in factual correctness which is good comparing to the existing studies
- The best performing RAG system was build using Reranking and Corrective RAG architecture

Links

- Project code: https://github.com/antoshsha/thesis_project
- Evaluation dataset: https://github.com/antoshsha/thesis_project/blob/main/questions.json

Thanks for your attention

Q&A