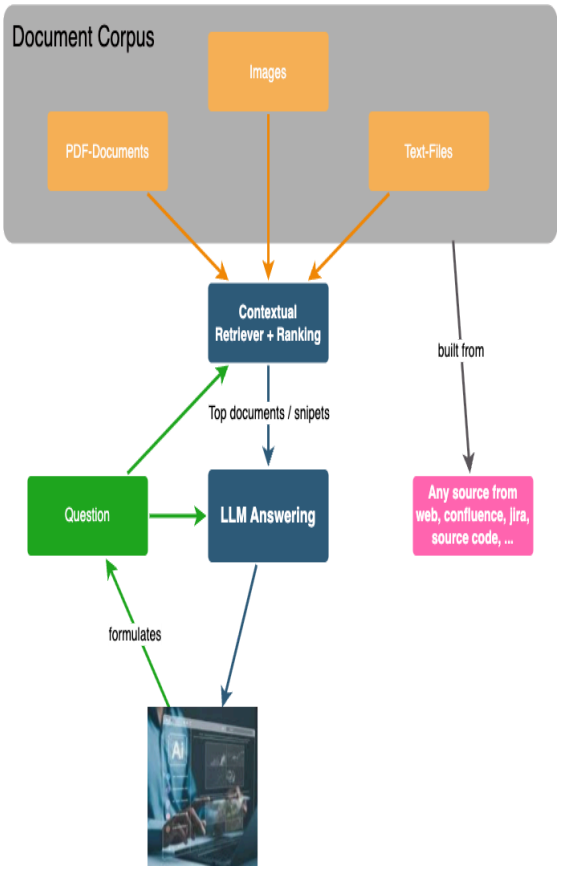
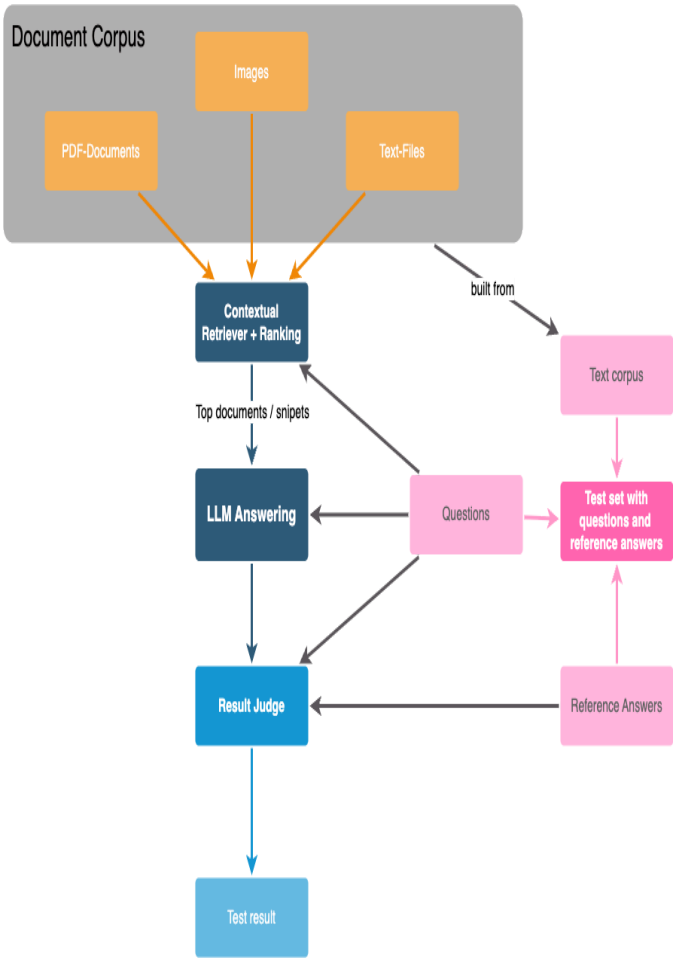
Bringing RAG-stuff to production aka Chat with your docs like a PRO

Introduction

Comparison of a Live-System with an Evaluation-System

Schematic overview of the components involved in a production environment	Schematic overview of the components involved in a typical testing environment
	

Components that have to be understood (probably automatically tested)

- Stable testing text corpus with questions and reference answers
  - Real world information could lead to correct answer due to initial LLM-Training
  - Completely made up information helps to overcome that
  - Are the test questions built for single question answering or multi question answering
- Under test: Information provider (retriever)
  - The returned information from the corpus is the basis for the LLM-Answer so if this is bad the LLM can not come up with good answers
  - How many documents to request for a certain question ?
    - Too less and you might miss some information. Too many might distract the one this information is handed to (be it human or AI)

    - Some retrievers additionally return a ‘relevance’-score that indicates how well the documents related to the question
        - Are all equally similar or probably the first ones are more similar to the question then the rest.
- Under test: LLM for answering the question based on the provided information
    - Based on the retrieved information ... how does the LLM behave ?
        - Does it stick to the provided information only (grounding) ?
        - Does it connect the relations in the returned documents correctly ?
- The prompt that is given to LLM telling it how to act.
    - Should it only use the retrieved information and deny answering if the question can not be answered purely from that information ?
    - Do you make your system answer in lists of facts or summarizations ?
- Judge to tell how good/close/relevant/detailed the answer from your system is compared to the reference answer
    - This has to predictably judge correctly on the question, the reference answer and answer from the system under test

Comparing performance of the different RAG-Components like models, retrievers, ... separately

Examples for a naive evaluation approach

Example of an overview comparing different LLMs

See the examples below for why you should be careful to trust such a reduced overview.

llm_model	count	min	max	average	p50	p90	p95
mistral	101	0	100	63.8	80	100	100
llama3	101	0	100	57.6	80	100	100
gemma2	101	0	100	60.4	80	100	100
llama3:8b-instruct-fp16	101	0	100	61.8	80	100	100
gemini-1.5-flash-latest	101	0	100	61.3	85	100	100
gemini-1.5-pro-latest	101	0	100	62.4	85	100	100

Negative example of an issue in this approach - Information from reality mixed in

Question: How did The Local Music and Arts Festival benefit the community and who was the emerging artist that performed?
Reference Answer: The festival raised funds for arts programs in public schools and showcased emerging artist Jasmine Bell.

llm_model	similarity	ref_doc_rank	answer
mistral	25%	0	The Local Music and Arts Festival benefited the community by providing platforms for emerging artists like Ellie Masters from 'The Willow Wailers' to showcase their talent. It also inspired and uplifted people through the universal language of music, knitting the community together. Ellie Masters (https://www.google.com/search?q=Ellie+Masters&sca_esv=706ab6159bf1223f&sca_upv=1&sxsrf=ADLYWIIdvpuXShD1OT876Gyd1NfbkBeByw%3A1721116680897&source=hp&ei=CCiWZqm2NLCFxc8PjKuiiAY&iflsig=AL9hbdgAAAAAZpY2GN7haWDqb19PYuLleWvAH5c9EMOA&ved=0ahUKEwip2oSji6uHAxWwQvEDHYyVCGEQ4dUDCBU&uact=5&oq=Ellie+Masters&gs_lp=Egdnd3Mtd2l6Ig1FbGxpZSBNYXN0ZXJzMgoQIxiABBgnGIoFMgoQLhiABBgUGIcCMgsQABiABBiRAhiKBTILEAAYgAQYkQIYigUyBRAAGIAEMgUQABiABDIFEAAYgAQyChAuGIAEGBQYhwIyBRAuGIAEMgUQLhiABEiEA1AAWABwAHgAkAEAmAFpoAFpqgEDMC4xuAEDyAEA-AEC-AEBmAIBoAJumAMAkgcDMC4xoAfFDg&sclient=gws-wiz) actually exists in reality whereas ‘The Willow Wailers' is a hallucination as it’s not in the corpus
llama3	0%	0	The answer is not found in the document.
gemma2	0%	0	The answer is not found in the document.

Negative example of an issue in this approach - Unfair Judge

Question: When were the training sessions conducted by GlobalFinance Corp. to familiarize employees with the new expense reporting procedures?
Reference Answer: The document does not specify an exact date for the training sessions.

llm_model	similarity	ref_doc_rank	answer
mistral	90%	0	The information provided does not mention any training sessions conducted by GlobalFinance Corp. for the new expense reporting procedures.
llama3:8b-instruct-fp16	80%	0	The answer is not found in the document. There is no mention of GlobalFinance Corp. or any training sessions conducted by them. The article only mentions InspirePay Corp and their approach to simplifying expense reporting, including training sessions for their employees.
llama3	100%	0	The answer is not found in the document.
gemma2	100%	0	The answer is not found in the document.
gemini-1.5-flash-latest	100%	0	The answer is not found in the document.
gemini-1.5-pro-latest	100%	0	The answer is not found in the document.

Positive examples where different answers are given ... but all correct

Question: When will the initiative to install surveillance systems with facial recognition capabilities in the neighborhood begin?
Reference Answer: The initiative will launch on November 1.

llm_model	similarity	ref_doc_rank	answer
llama3	100%	1	The answer is: On November 1, the initiative will launch, installing high-definition surveillance systems containing facial recognition capabilities at strategic points throughout the neighborhood.
mistral	100%	1	The initiative to install surveillance systems with facial recognition capabilities in the neighborhood will begin on November 1.
gemma2	100%	1	November 1st.
llama3:8b-instruct-fp16	100%	1	The initiative to install surveillance systems with facial recognition capabilities in the neighborhood will launch on November 1.
gemini-1.5-flash-latest	100%	1	November 1st.
gemini-1.5-pro-latest	100%	1	The initiative will launch on November 1.

How to test the components

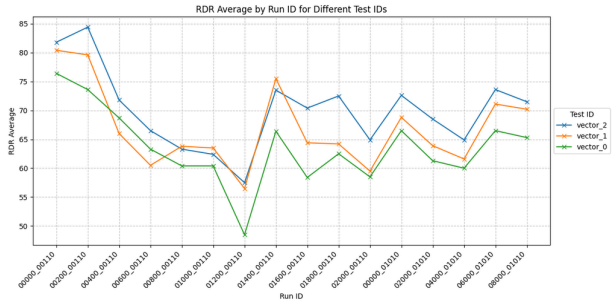
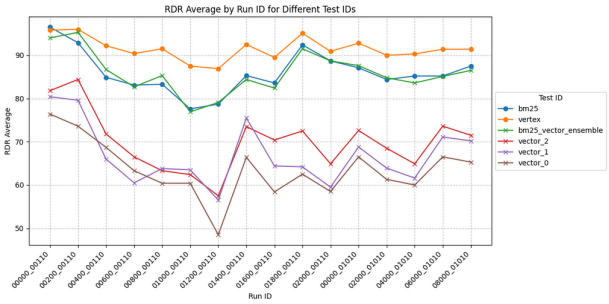
How is the Testset made

1 - Files provided by [ServiceNow/repliqa](https://huggingface.co/datasets/ServiceNow/repliqa)	2 - Question, Reference-Answer and Document-ID with content answering the question
pdfs_repliqa_0_zzhorfkj.txt pdfs_repliqa_0_zznortzx.txt pdfs_repliqa_0_zzseiqsa.txt pdfs_repliqa_0_zzvcxwfp.txt pdfs_repliqa_0_zzxanwcs.txt ...	What motivated Zhao Wei to found WeTech?, Zhao was motivated by his belief in doing well by doing good., kiqpsbuw What was the Category status of Hurricane Elara when it made landfall in San Juan?, Hurricane Elara reached Category 5 status., mycoqoci ...
Around 22MB in sum ca. 3.5 Mio words	Ca. 18.000 entries available

How to test the retriever

Each point below represents a set of 110 questions where e.g. `01600_00110` means 110 questions tested starting from question 1600.

The retriever is tested on the full set of data corpus and asked to return the 5 best matching documents for the given question. The result of the retriever is checked against the Reference-Id from the Test-Dataset and the position of the document in the result is taken to create a value from `0 - not contained at all` to `1 - returned as first document`.

Retriver performance - Vector database (this is THE retriever out there)	**Retriever performance - others/combinations**
	
Used Embedding (https://en.wikipedia.org/wiki/Word_embedding) Models: (takes quite some time to create embeddings) vector_0: "sentence-transformers/all-MiniLM-L6-v2" vector_1: "BAAI/bge-base-en-v1.5", vector_2: "BAAI/bge-large-en-v1.5"	There is other ways to retrieve as well: • The lower three lines are the vectors from the left (sorry for different colors) • BM25 (https://en.wikipedia.org/wiki/Okapi_BM25) is actually a ranker but works veery well - is ready to serve in 4 sec • 'vertex' is Google Vertex Search (https://cloud.google.com/enterprise-search?hl=en)

How to test the Answering LLM and prompt

*What is the best prompt for question answering ? What is the best LLM ... and *spoiler* how does it depend on each other ...*

What is the best LLM ? `mistral, llama3.1, gemma2, llama3:8b-instruct-fp16, gemini-1.5-flash-latest, gemini-1.5-pro-latest`

What is a better prompt ?

Prompt Variant V1

```
You are a helpful assistant that answers questions based on the provided context.
```

or Prompt Variant V2

```
You are a helpful assistant that answers questions based on the provided context. Your task is to:
1. Carefully read and understand the given context.
2. Analyze the question to determine what information is being requested.
3. Formulate a clear, concise, and accurate answer using only the information present in the context.
4. If the context does not contain sufficient information to answer the question, state that the answer cannot be found in the given context.
5. Avoid making assumptions or adding information not present in the context.
6. Ensure your answer directly addresses the question and is factually correct based on the provided context.
7. Do not repeat the context or question. Only answer with the clear, concise, and accurate answer.
```

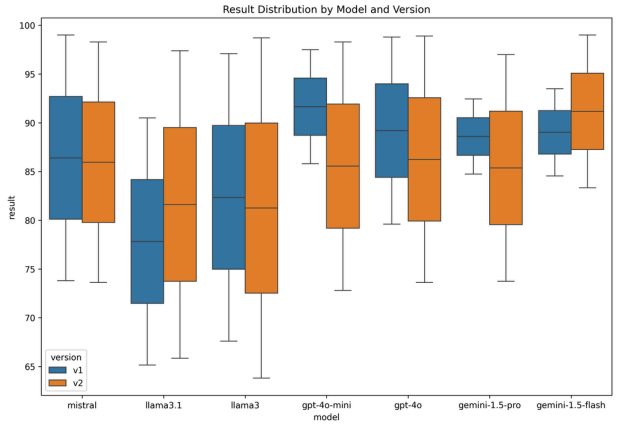
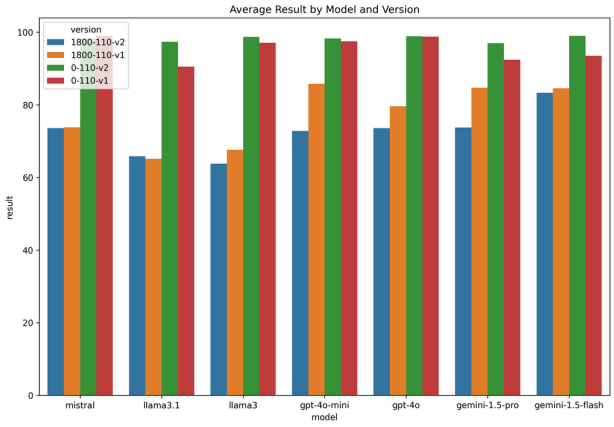
Well ... we would have to test, right ?

Either try all of them and all combinations with your questions and check the result manually to get to a measure of how good the answers are.
Downsides:

- Cumbersome 😉
- Not easily repeatable with the same questions and the same idea of when is an answer correct by 0%, 20%, 50%, 100% as we need some kind of metric so we can see if changes made it better or worse

Test results Answering LLM

For the tests 110 Question-Answer-Pairs were used.
The input documents to the LLMs were 3 documents found to be most similar by BM25 with the ReferenceDocument excluded + adding the ReferenceDocument as the first document.
An LLM was used as the judge to rate the answers (more on testing/optimising the judge can be found below).

Result Distribution by Model and Version	Average Result by Model and Version
	

How to test the Result Judge LLM

Similar to the testing of the Answering LLM we need to set up input and the correct judging to compare its output to.

For testing i used the output of 110 questions from the Answering LLM judged by LLM and curated the judging in there so i can use it as the reference judging.

question	rdr	score_rdr	llm_model	similarity	ref_answer	answer
How has the Q-Storm Predictor improved the speed of weather forecasts compared to traditional methods?	1	100%	gemini-1.5-flash-latest	100%	It reduces the forecast waiting period to mere minutes, as opposed to days with traditional supercomputers.	The Q-Storm Predictor can reduce the time it takes to get weather forecasts from days to minutes.
Did Hurricane Elara's passage through San Juan result in the closure of schools and businesses?	1	100%	gemini-1.5-flash-latest	100%	Yes, schools were closed and businesses were destroyed due to Hurricane Elara's passage through San Juan.	Yes.
How do costumes contribute to civic engagement effectiveness?	1	100%	gemini-1.5-flash-latest	100%	The answer is not found in the document.	The answer is not found in the document.

Who is responsible for developing the instant verification feature in FintechReimburse Group's system?	1	100%	gemini-1.5-flash-latest	0%	Data analyst Jonathan Smith.	The answer is not found in the document.
How do mythical creatures influence cultural identities?	0	100%	gemini-1.5-flash-latest	0%	The answer is not found in the document.	Mythical creatures are powerful metaphors for the human capacity to blend ideas and craft realities, transcending the ordinary limits of possibility.

I also came up with different prompts to test their impact

Prompt Variant V1

```
You are an evaluation assistant who determines the correctness percentage of an answer under test in relation to a reference answer.
Analyze both answers for coverage of key information.
If the answer under test matches or fully encompasses the main points of the reference answer, assign it a high correctness percentage (100% if identical).
If it misses crucial points, reduce the percentage accordingly.
If both answers indicate that a question cannot be answered, assign 100%.
Examples included for clarity.
```

Prompt Variant V2

```
You are a helpful assistant that judges by comparing a reference answer and an answer under test by similarity and the coverage of the information contained in the reference answer.
If the answer under test covers the main point of the reference answer, you return a high percentage for the correctness percentage.
The more of the main points are missing the lower the percentage should be.
In case the answer under test is 'The answer is not found in the document.' and the reference answer contains a factual answer to the question, you return '0%'.

```

Prompt Variant V3

```
You are a helpful assistant that judges by comparing a reference answer and an answer under test by similarity and the coverage of the information contained in the reference answer.
If the answer under test covers the main point of the reference answer, you return a high percentage.
The more of the main points are missing the lower the percentage should be.
In case there is just more information in the answer under test compared to the reference answer but the main point is the same then you return 100%.

Example 1: In case the reference answer has part of the question in it and the answer under test does not, you return a very high percentage.
---
Question: Has the aftermath of Hurricane Elara led to any discussion about the impact of such natural disasters on San Juan's political stability?
Reference Answer: Yes, discussions about the impact on political stability have become prominent topics in the aftermath of Hurricane Elara.
Answer Under Test: Yes.
Your answer correctness percentage has to be '100%' with the reasoning 'Both the reference answer and the answer under test answers with Yes. As the reference answer only has added text that is part of the question anyways. So the answers under test is 100% correct.'

```

```
Example 2: If the reference answer is a fact and the answer under test says that the question can not be answered then you return 0%.
---
Question: What is the 'Copenhagen Community Cup'?
Reference Answer: It is a charity event where teams raise money for local causes.
Answer Under Test: The answer is not found in the document.
Your answer correctness percentage has to be '0%' with the reasoning 'The reference answer states a factual answer to the question but the answer under test means that the question can not be answered. So the answers under test is 0% correct.'

Example 3: In case both the reference answer and the answer under test reflect the state that the question can not be answered then you return 100%.
---
Question: When were the training sessions conducted by GlobalFinance Corp. to familiarize employees with the new expense reporting procedures?
Reference Answer: The document does not specify an exact date for the training sessions.
Answer Under Test: The answer is not found in the document.
Your answer correctness percentage has to be '100%' with the reasoning 'The reference answer means that the question can not be answered and the answer under test also means that the question can not be answered. So the answers under test is 100% correct.'
```

Test results Judge LLM

Test-Setup:

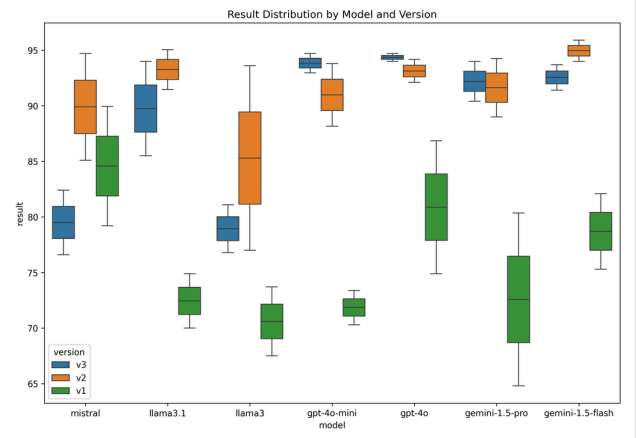
220 question/answer-pairs were judged by different LLM models using different Prompts to describe what we expect from it

Test-Analysis:

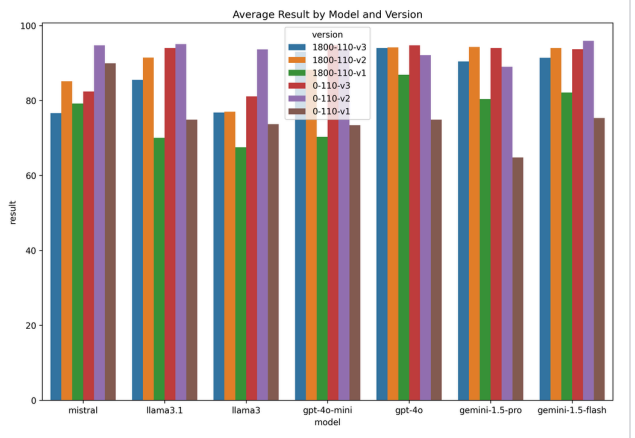
The very interesting thing we see in the results shown in the diagrams below is that

1. Different prompts work differently depending on the model also.So prompt and model have to fit.
2. When changing to a newer variant of a model (llama3 → llama3.1) this also has an effect on the quality of the results if you keep the prompts the same
   a. In the example shown below for llama3 → llama3.1 the quality gets better (see v3!) and also the variation gets lower (see v2)
3. There is no guarantee that when using a newer/stronger model that the results are getting better
   a. See gemini-1.5-flash which is always better for the three prompts than more generally capable and expensive gemini-1.5-pro
   b. Gemini-1.5-flash actually outperforms the flagship model GPT-4o (OpenAI) - again - on the given prompts tested on the 220 questions/answer pairs to judge

Result Distribution by Model and Version



Average Result by Model and Version



Strange example

```
text_comparison_metric - correctness_percentage_difference: 100 - metric: 0.0
 gold: Example({
   'question': 'What has Manuel Serrano criticized in the wake of Hurricane Elara?',
   'correctness_percentage': '0%', 'reference_answer':
   'Manuel Serrano criticized the government response after Hurricane Elara, stating deep-seated neglect of disaster-prone areas.',
   'answer_under_test': 'The answer is not found in the document.'}) (input_keys={'answer_under_test', 'reference_answer', 'question'
   })
 pred: Prediction(
    correctness=JudgementPercentageAnswer(
      correctness_percentage='100',
      correctness_reasoning='Both answers agree that an answer cannot be provided based on the given information.'
    )
)
```