GPTx und RAG in der Praxis

Schluss mit Prototyp

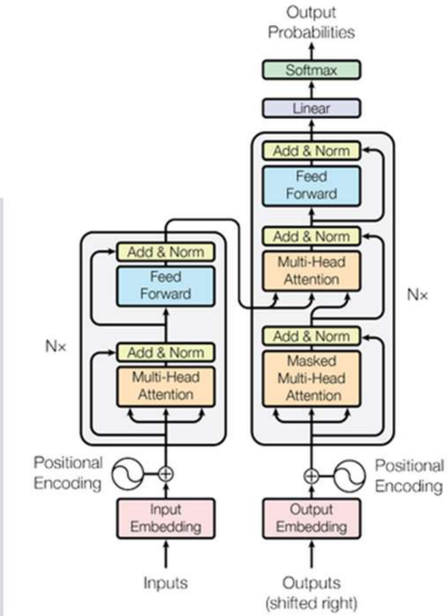
Olliver Zeigermann
Christian Hidber

data2day, Heidelberg, September 2024

Chef:
mein Enkel kann das auch...

LLM Intro

Transformers, LLMs, Encoder, Decoder: WTF?



- **Transformers**: A flexible architecture that uses self-attention to process sequential data efficiently.
- **LLMs**: Large-scale Transformer models trained on extensive text datasets to perform various language tasks.
  - **Encoder Models**:
    - Part of the Transformer architecture focused on understanding and interpreting input data (*e.g. BERT*)
    - Instrumental for Embedding Models
  - **Decoder Models**:
    - Part of the Transformer architecture focused on generating sequential output based on the interpreted inputs or prior outputs
    - Instrumental for GPT-style Models like **Llama, Mistral or OpenAI GPT**

Decoder Models

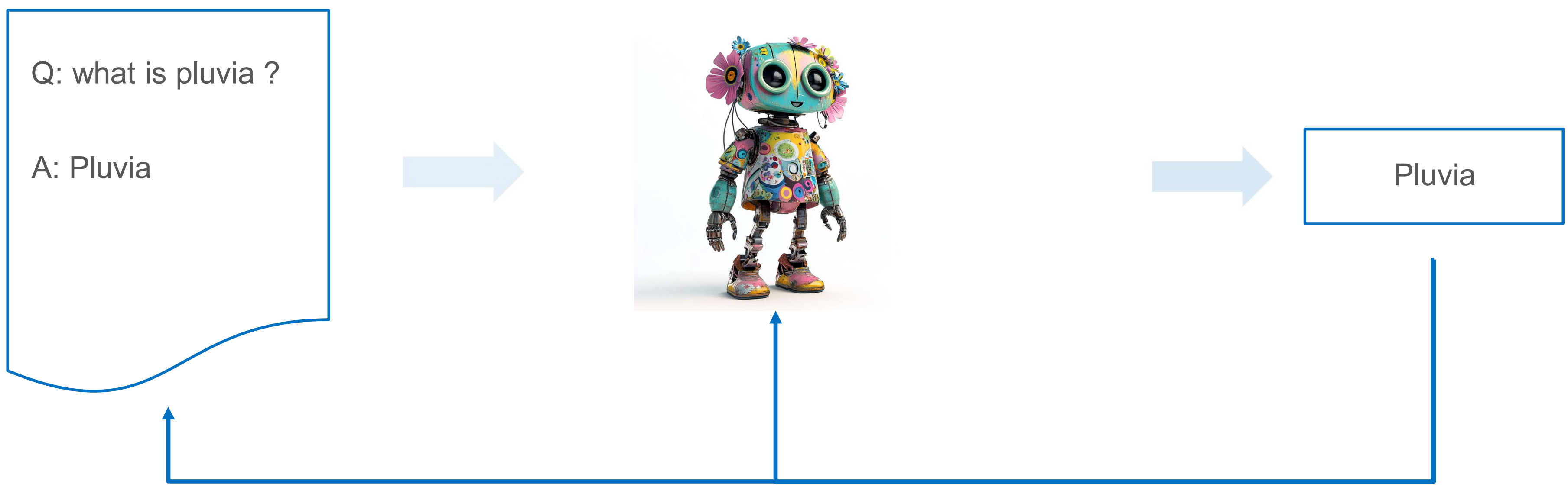
How does a Decoder Model work ?

Q: what is pluvia ?

- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- **«the context»**

- Trained on huge datasets
- Does not change
- Same for all users
- **«the model»**

How does a Decoder Model work ?

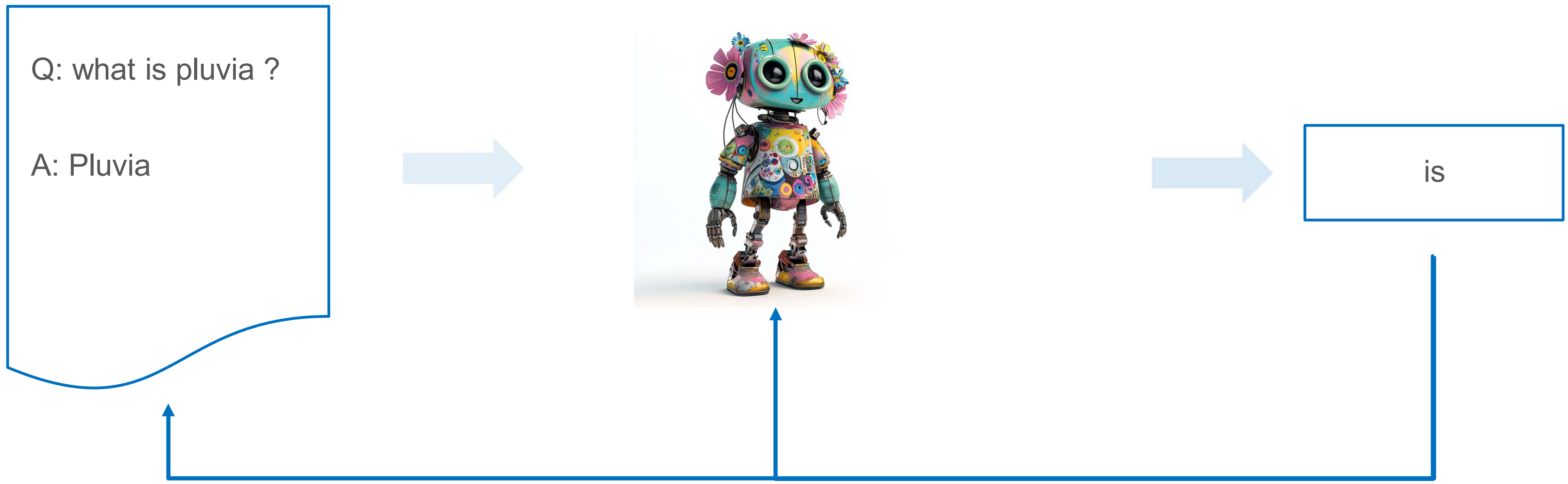


- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- **«the context»**

- Trained on huge datasets
- Does not change
- Same for all users
- **«the model»**

- Single «word»
- Depends on context and model
- **«the token»**

How does a Decoder Model work ?

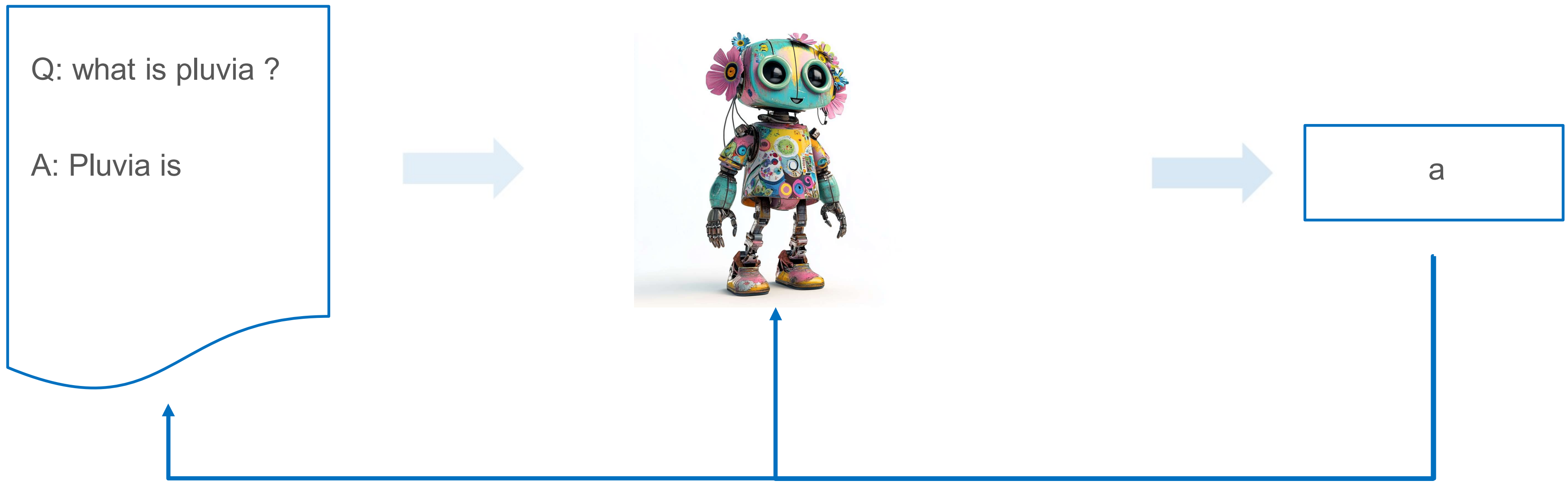


- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- **«the context»**

- Trained on huge datasets
- Does not change
- Same for all users
- **«the model»**

- Single «word»
- Depends on context and model
- **«the token»**

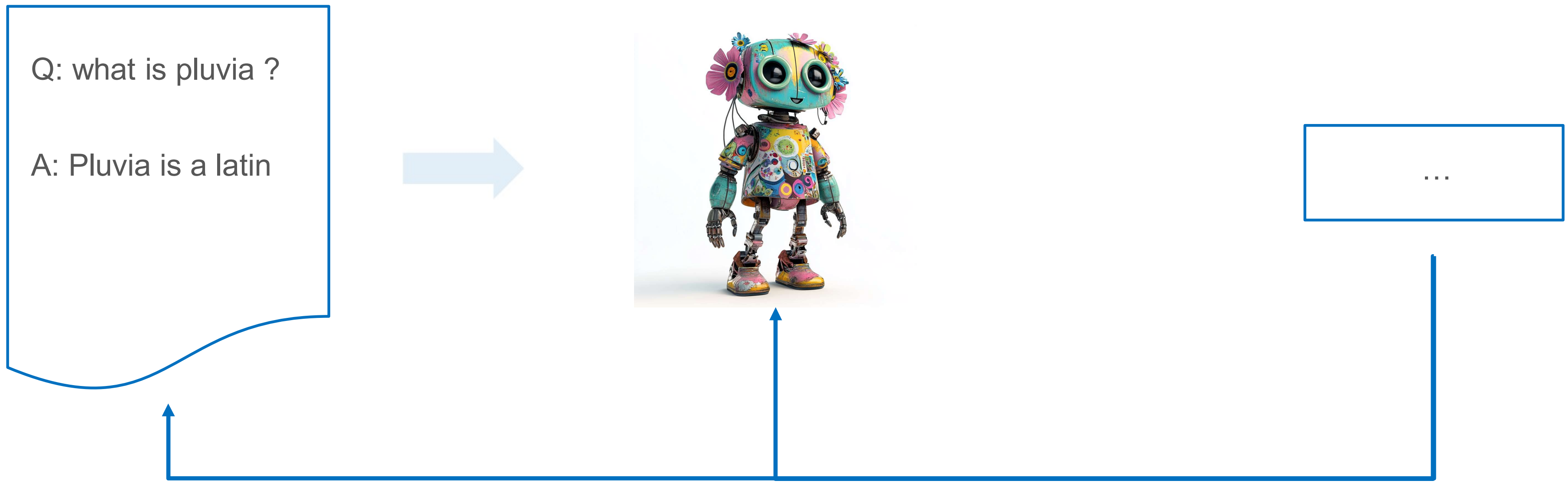
How does a Decoder Model work ?



- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- «the context»

- Trained on huge datasets
- Does not change
- Same for all users
- «the model»

- Single «word»
- Depends on context and model
- «the token»



- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- «the context»

- Trained on huge datasets
- Does not change
- Same for all users
- «the model»

- Single «word»
- Depends on context and model
- «the token»

How does a Decoder Model work ?

Q: what is pluvia ?

A: Pluvia is a latin word meaning rainfall.

EOT

- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- «the context»

- Trained on huge datasets
- Does not change
- Same for all users
- «the model»

- Single «word»
- Depends on context and model
- «the token»

Naïve Approach

You are an expert in …..

What is Pluvia ?

Use the following facts:

User:
Asking a Question

answer

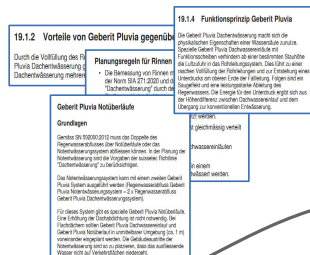
Llm:
Generating an Answer

Idea: just a few pages

You are an expert in …..

What is Pluvia ?

Use the following facts:



User:
Asking a Question

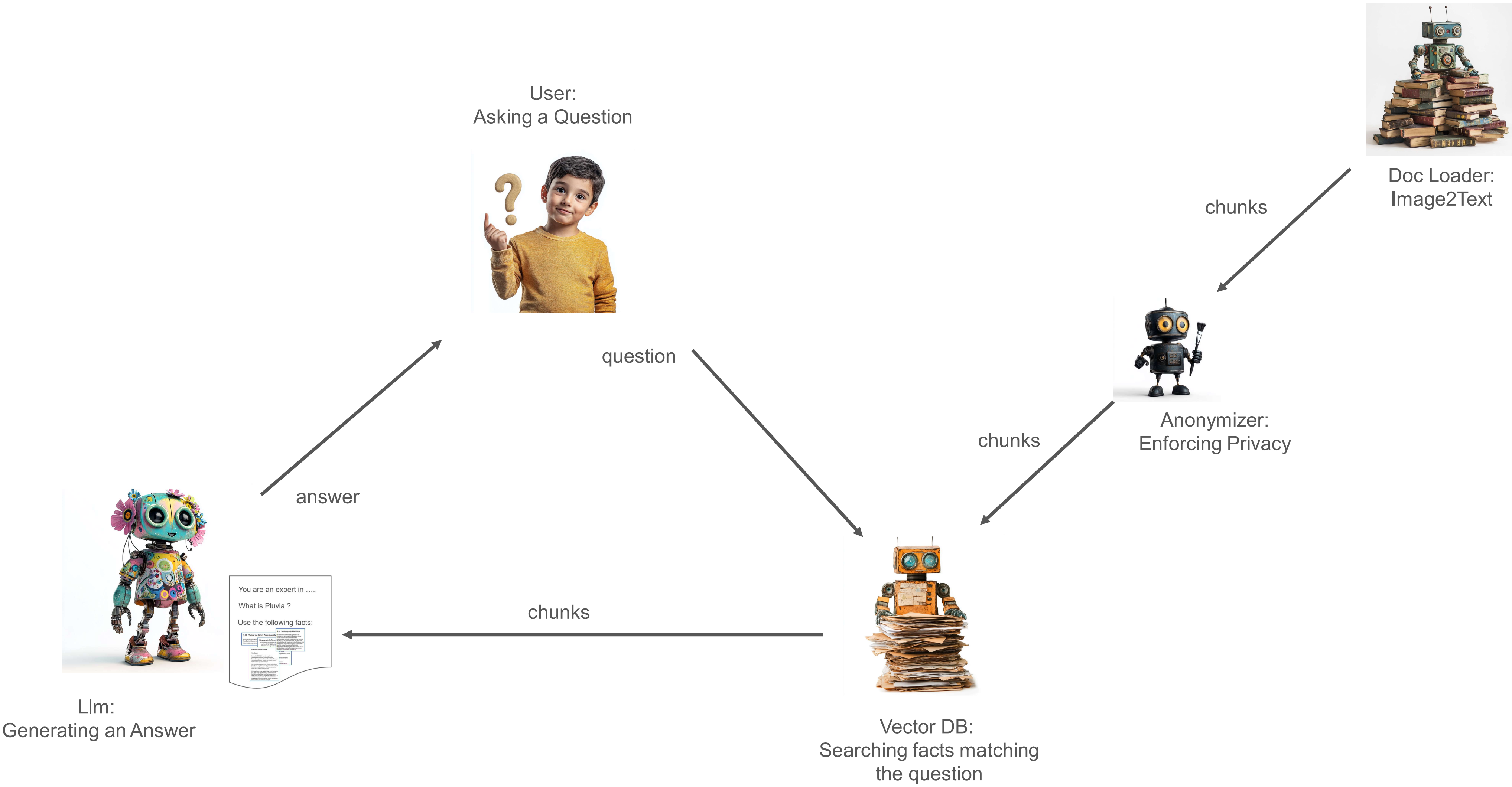
answer

Llm:
Generating an Answer

RAG

Retrieval Augmented Generation

RAG System Architecture



Demo:
Low Risk RAG Applications

Choosing an application

Low Risk, but nice benefit

- What is the biggest risk we are facing?
- What is the worst thing that could happen and how to mitigate that?

- Choose something that is low risk, but nice benefit
- Low profile
- Failures should be ok
- Let the whole organization learn
- Management likes it, but is afraid

From Prompt Hacking to Production

Der kleine Heimwerker vs Ingenieur Tätigkeit

- hohes Maß an Automatisierung
- Generalisierung: Modell auf zukünftige Daten verallgemeinert anwendbar?
  - Der Nutzen ergibt sich durch Vorhersage auf bisher unbekannten Daten in der Zukunft.
  - Auf diesen muss das Modell eine gute Leistung bringen. Nur das ist relevant.
- über einen längeren Zeitraum stabil bleiben
- **Beispiel LLM und Prompting - können wir das nicht alle?**
- Von wqelcher Art Prompting sprechen wir ? Muss man unterscheiden von ad hoc Prompting
- Bei Ad hoc sieht man direkt, ob es geht. Man hat ein hohes Maß an menschlicher Überwachung.
- Unterschied internes Werkzeug und stabiler Service. Plattform Team

Evaluation

Evaluation on text results

User:
Asking a Question

answer

Llm:
Generating an Answer

Human Eval

Question

- What is Pluvia ?

Answer

- Pluvia is a latin word meaning rainfall.
- The latin word for rainfall.
- ….

=> equality not an option

Evaluation on text results

User:
Asking a Question

answer

Llm:
Generating an Answer

Evaluation Criteria:

- Correct
- Complete
- Concise
- Relevant
- Contradiction free
- Language
- Style
- …
- Generation successful

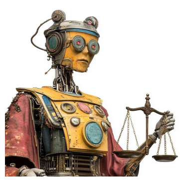
Ground Truth

Human Eval

Statistics

Evaluation on text results

User:
Asking a Question



answer

Llm:
Generating an Anwer

LLM as a Judge

Ground Truth

Human Eval

Statistics

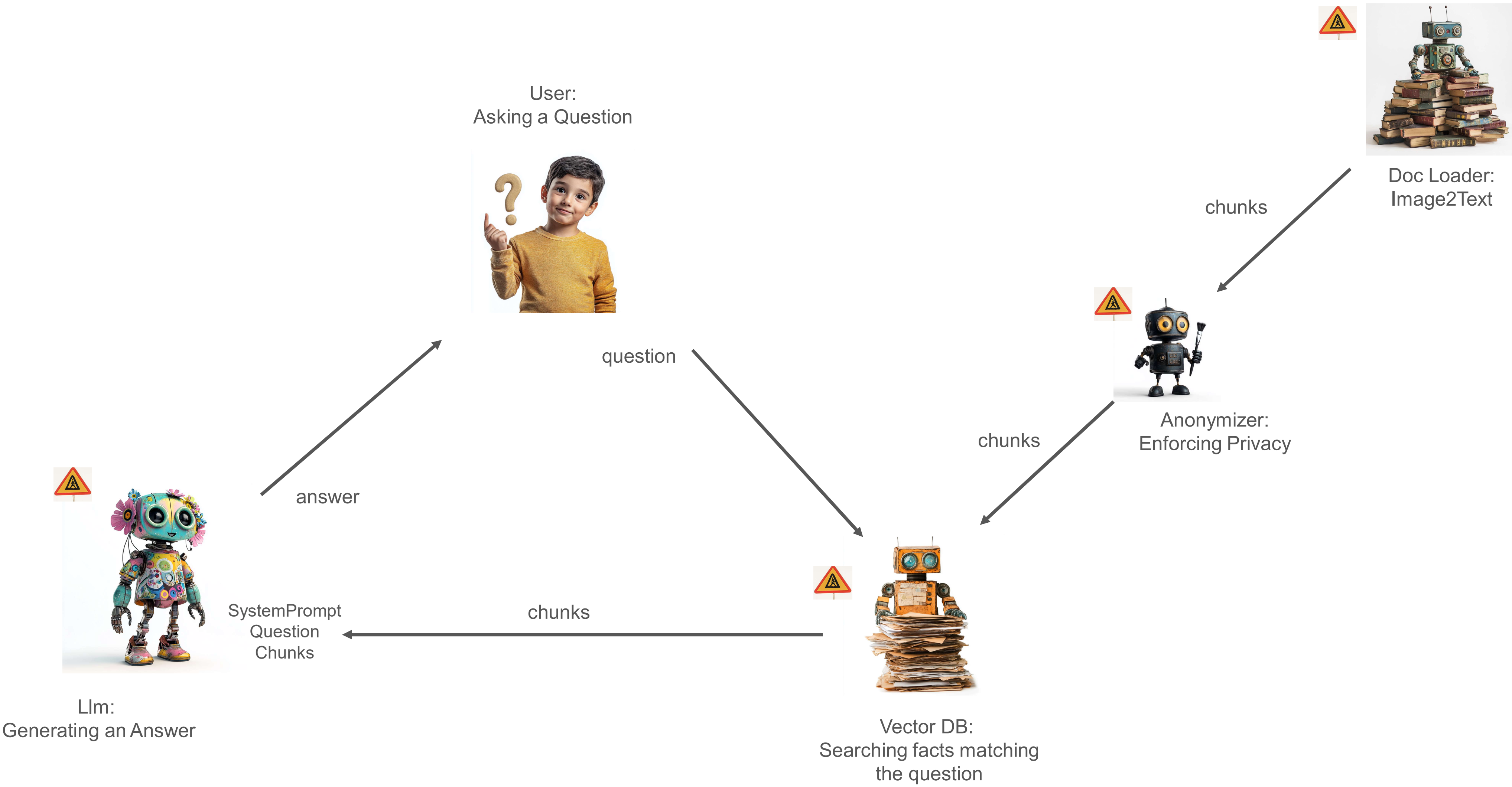
Evaluation Criteria:

- Correct
- Complete
- Concise
- Relevant
- Contradiction free
- Language
- Style
- …
- Generation successful

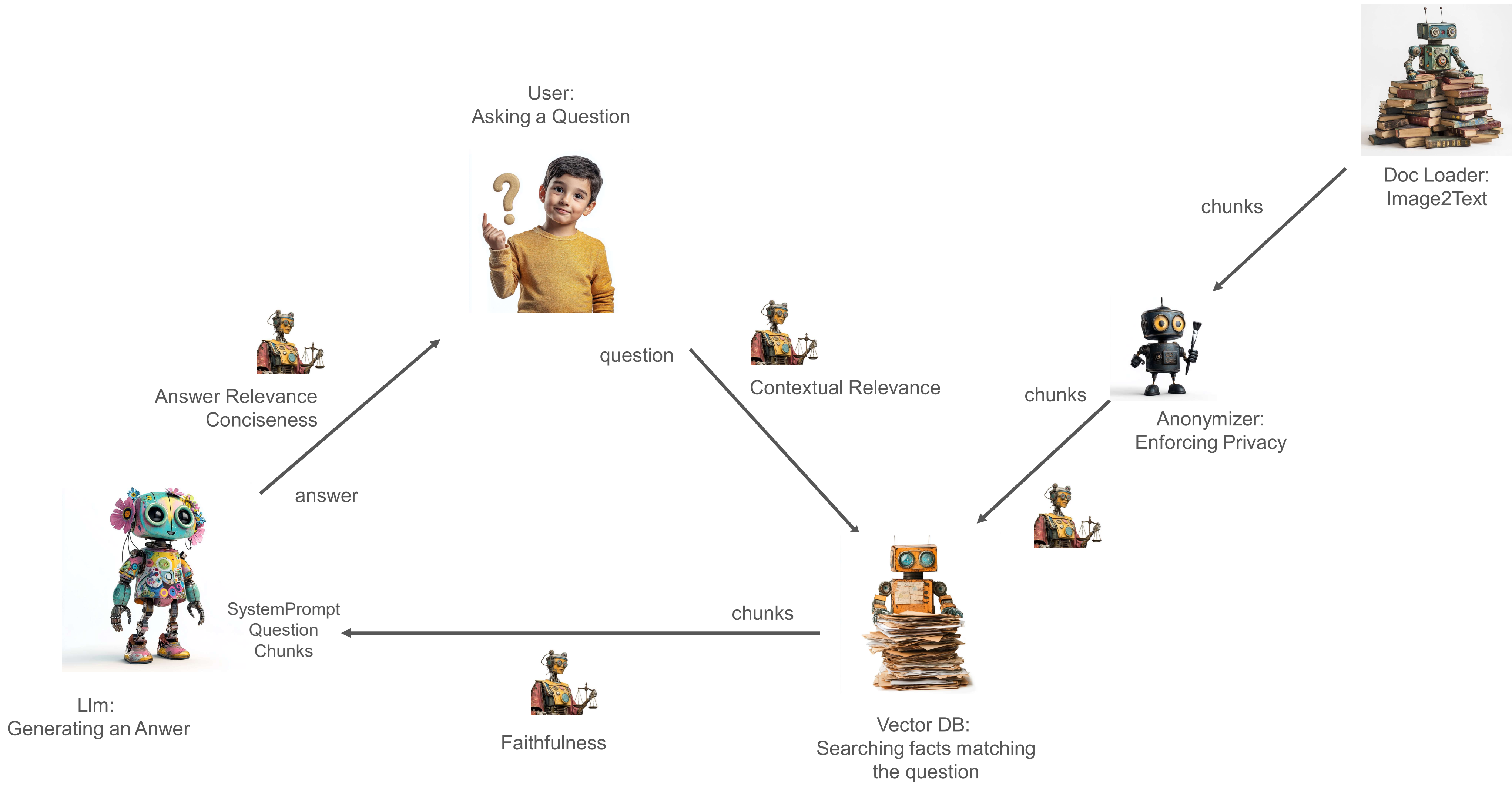
Demo:
Evaluation Notebook

https://colab.research.google.com/github/DJCordhose/llm-from-prototype-to-production/blob/main/Eval4pptx.ipynb

RAG System Architecture



RAG System Architecture: Online Evaluation



Online Eval: Example


```
[07:56:37 INF] POST "https://[redacted].azurewebsites.net/"eval succeeded in 4.91s with response={
  "MetaData": {
    "Answer": "Der Artikel Sigma20 BetÃ¤tigungsplatte dient zur Steuerung der 2-Mengen-SpÃ¼lung bei Geberit UP-SpÃ¼lkÃ¤sten. Sie ermÃ¶glicht die Auswahl zwischen einer groÃŸen und einer kleinen SpÃ¼lmenge, um W
    "CreatedDate": "2024-09-10T07:56:37.744166Z",
    "DeepEval": {
      "Answer_Relevancy": {
        "reason": "The score is 1.00 because the response directly addresses the purpose of the Sigma20 BetPl. article without any irrelevant statements.",
        "score": 1.0
      },
      "Conciseness_(GEval)": {
        "reason": "The output is somewhat concise but includes unnecessary details about materials and suitability that could be omitted for brevity.",
        "score": 0.6
      },
      "Contextual_Relevancy": {
        "reason": "The score is 0.33 because the context discusses various models and specifications of flushing systems but does not provide any information about the article 'Sigma20 BetPl.' or its purpose.",
        "score": 0.3333333333333333
      },
      "Faithfulness": {
        "reason": "The score is 0.80 because the actual output inaccurately generalizes the material of the BetÃ¤tigungsplatte, stating it could be made of Edelstahl, while the retrieval context clarifies that
        "score": 0.8
      }
    },
    "ElapsedSeconds": 4.85,
    "EvalType": "deep_eval",
    "EvalVersion": "240903",
    "Input": "Wozu dient der Artikel Sigma20 BetPl., fÃ¼r 2-Mengen-SpÃ¼lung weiÃŸ / weiÃŸ matt ?"
  },
  "Metrics": {
    "Answer_Relevancy": 1.0,
    "Conciseness_(GEval)": 0.6,
    "Contextual_Relevancy": 0.3333333333333333,
    "Faithfulness": 0.8
  },
  "Score": 0.6833333333333333
}
```

Online Eval: Example

```
LlmDesc germany     W.240820_C.240625         : Score=0.000 TextGenerated=0.000  [counts 242,242]
LlmDesc germany     W.240820_C.240625_E.240903: Answer_Relevancy=0.962 Conciseness_(GEval)=0.628 Contextual_Relevancy=0.341 Faithfulness=0.890 Score=0.705 TextGenerated=1.000
LlmDesc switzerland W.240820_C.240625         : Score=0.000 TextGenerated=0.000  [counts 140,140]
LlmDesc switzerland W.240820_C.240625_E.240903: Answer_Relevancy=0.943 Conciseness_(GEval)=0.613 Contextual_Relevancy=0.534 Faithfulness=0.851 Score=0.735 TextGenerated=1.000
LlmFp   germany     W.240820_C.240625         : Score=0.000 TextGenerated=0.000  [counts 4,4]
LlmFp   germany     W.240820_C.240625_E.240903: Answer_Relevancy=0.964 Conciseness_(GEval)=0.595 Contextual_Relevancy=0.739 Faithfulness=0.842 Score=0.785 TextGenerated=1.000
LlmFp   switzerland W.240820_C.240625         : Score=0.000 TextGenerated=0.000  [counts 8,8]
LlmFp   switzerland W.240820_C.240625_E.240903: Answer_Relevancy=0.984 Conciseness_(GEval)=0.624 Contextual_Relevancy=0.621 Faithfulness=0.757 Score=0.747 TextGenerated=1.000
LlmHelp german      W.240820_C.240625         : Score=0.000 TextGenerated=0.000  [counts 66,66]
LlmHelp german      W.240820_C.240625_E.240903: Answer_Relevancy=0.992 Conciseness_(GEval)=0.692 Contextual_Relevancy=0.731 Faithfulness=0.897 Score=0.828 TextGenerated=1.000
LlmSi   germany     W.240820_C.240625         : Score=0.000 TextGenerated=0.000  [counts 6,6]
LlmSi   germany     W.240820_C.240625_E.240903: Answer_Relevancy=0.987 Conciseness_(GEval)=0.615 Contextual_Relevancy=0.827 Faithfulness=0.864 Score=0.823 TextGenerated=1.000
LlmSi   switzerland W.240820_C.240625         : Score=0.000 TextGenerated=0.000  [counts 8,8]
LlmSi   switzerland W.240820_C.240625_E.240903: Answer_Relevancy=0.985 Conciseness_(GEval)=0.623 Contextual_Relevancy=0.904 Faithfulness=0.822 Score=0.833 TextGenerated=1.000
LlmTtp  english     W.240820_C.240625_E.240903: Answer_Relevancy=1.000 Conciseness_(GEval)=0.500 Contextual_Relevancy=0.000 Faithfulness=1.000 Score=0.625 TextGenerated=1.000
LlmTtp  french      W.240820_C.240625_E.240903: Answer_Relevancy=1.000 Conciseness_(GEval)=0.600 Contextual_Relevancy=0.000 Faithfulness=1.000 Score=0.650 TextGenerated=1.000
LlmTtp  german      W.240820_C.240625         : Score=0.000 TextGenerated=0.000  [counts 74,74]
LlmTtp  german      W.240820_C.240625_E.240903: Answer_Relevancy=0.925 Conciseness_(GEval)=0.611 Contextual_Relevancy=0.442 Faithfulness=0.899 Score=0.719 TextGenerated=1.000
[09:21:55 INF proplanner] HTTP POST /api/descriptions responded 200 in 25.1855 ms
```

Evaluation Issues

- Online Performance impact on LLM
  - Eval may call 10x more often, but have less output tokens
- Which LLM do you use ? Same ? Faster ? Most Powerful ?
- What Dimensions do you eval ?
  - Toxicity, Conciseness, Answer Relevance ?
  - Ground Truth available ?
- Human Feedback from your users ?
- Interpretation of the Scores ?

Eval Frameworks

- DeepEval https://docs.confident-ai.com/
- Ragas https://ragas.io/
- TruLens https://www.trulens.org/
- Evidently https://www.evidentlyai.com/
- Ares https://ares-ai.vercel.app/
- …

Your Experience ?

- Anyone doing RAG ? In Production ?
- Do you do evaluation ? By humans ?
- What else do you use for evaluation ?

Wrap Up

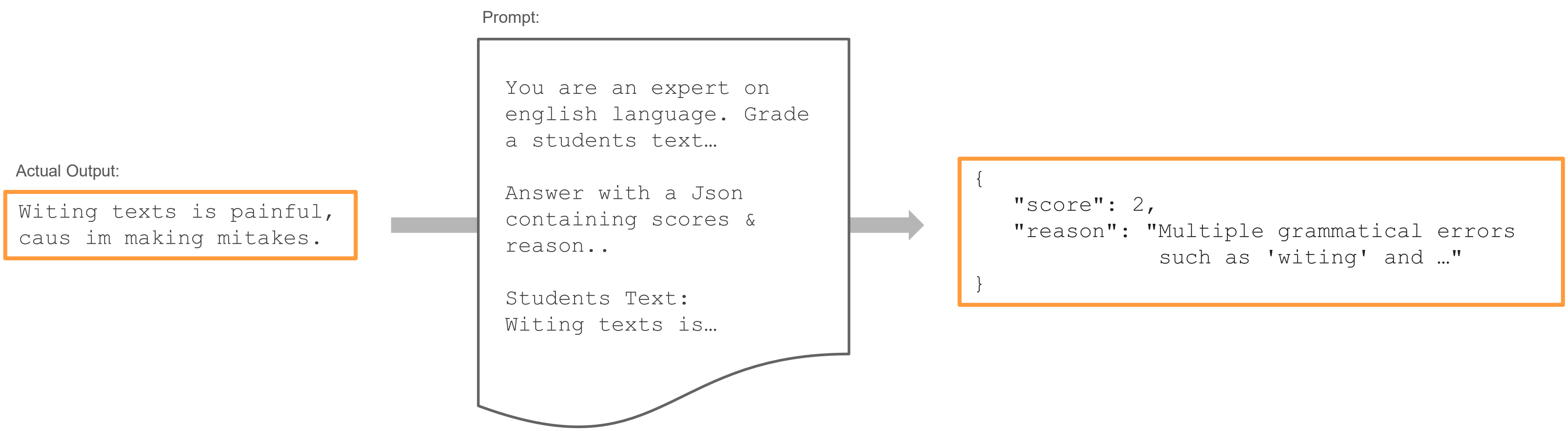
Key takeaways

- Human Eval is a great starting point
- LLM-as-a-Judge works, but take the scores with a grain of salt
- Use a strong LLM for evaluation
- Evaluation is even more crucial when using potentially less powerfull models
- Getting the Documents & keeping the up-to-date can be painful

Vorsicht vor dem Enkel des Chefs…

Thank you

Llm-as-a-judge: Idea



Llm-as-a-judge: G-Eval