Running LLMs on prem

A Practical Guide

Olliver Zeigermann
Christian Hidber

ODSC Europe, London, 9/5/24

Why self hosting an LLM?

You might want control over

1. Privacy / data protection
2. Availability and Scaling
3. Latency
4. Limitations
5. Cost of operation
6. Ecological footprint
7. Stability

Architecture Decision: self-hosting ?

- Decision : yes
- Key-Driver : **privacy**, version stability, limitations
- Challenge : which LLM ? Does it run on your hardware ?
- Surprises : operating GPUs is really hard
  - Failures
  - Power
  - Stability
  - Complexity, Work, Troubles

This workshop is about solving the challenges arising from self hosting

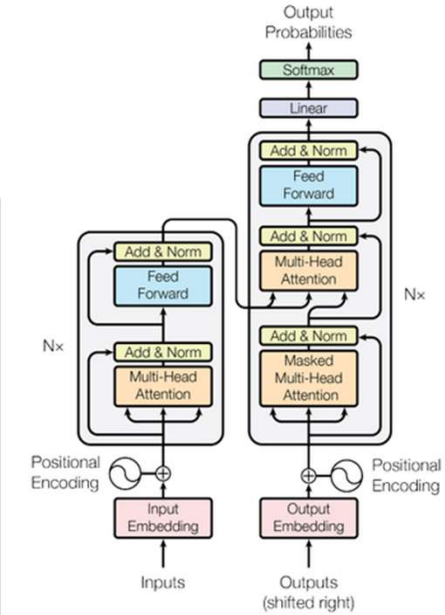
Who are we?

Olliver

Christian

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 ?

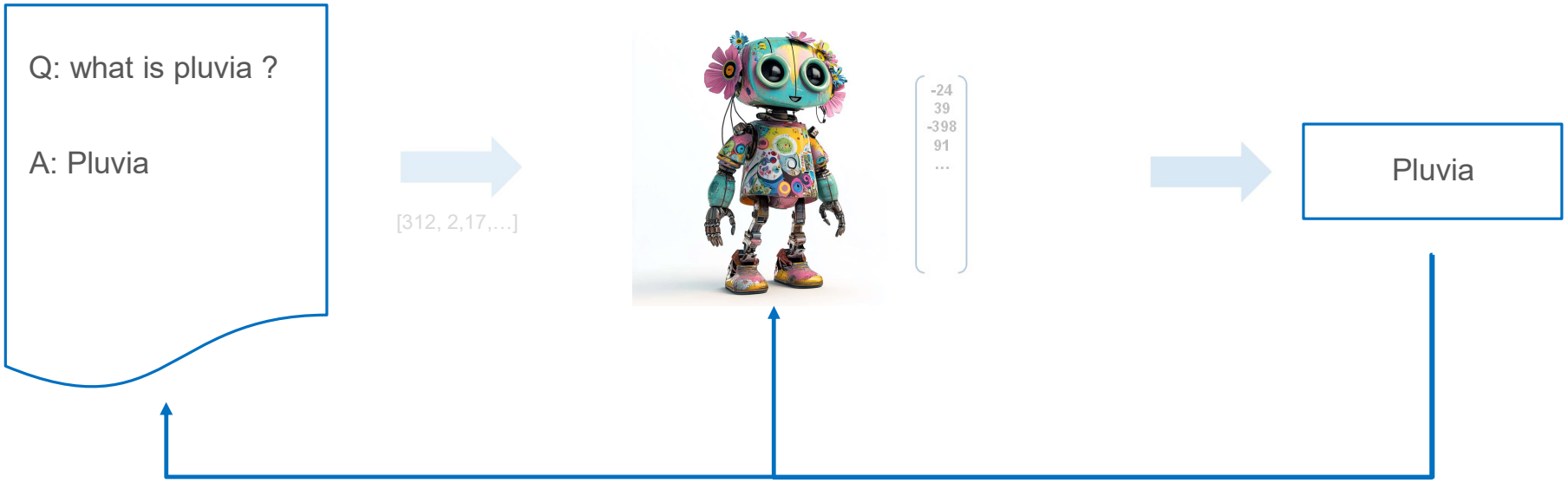
Q: what is pluvia ?

[312, 2,17,…]

- 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 ?

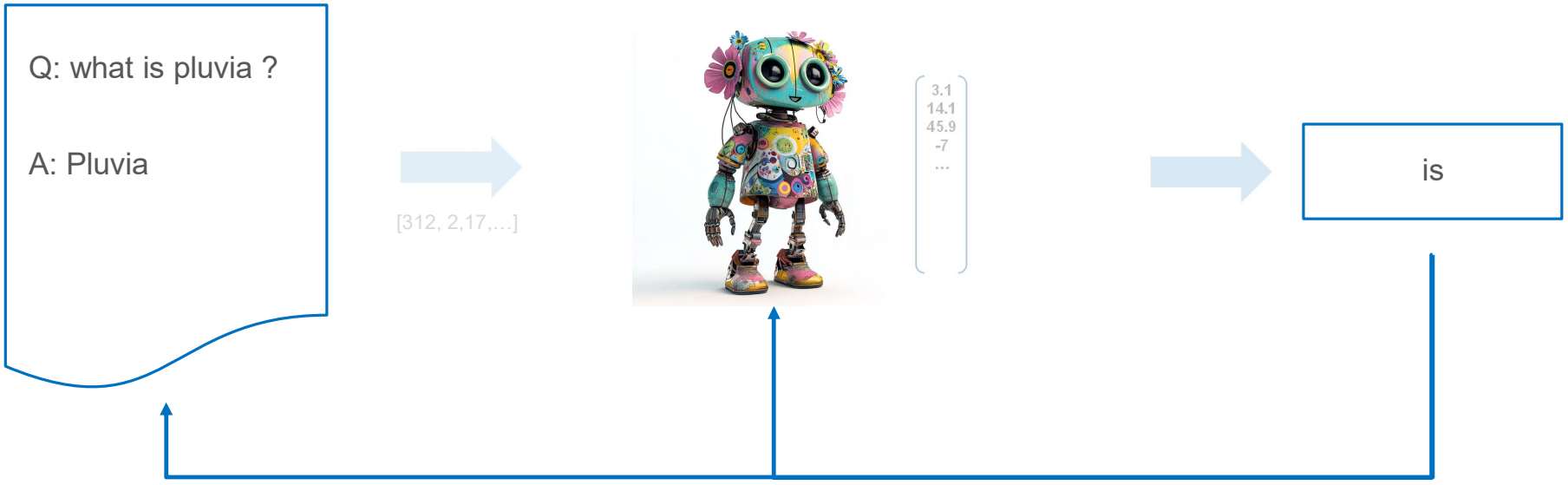


- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- **«the context»**

- 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 ?

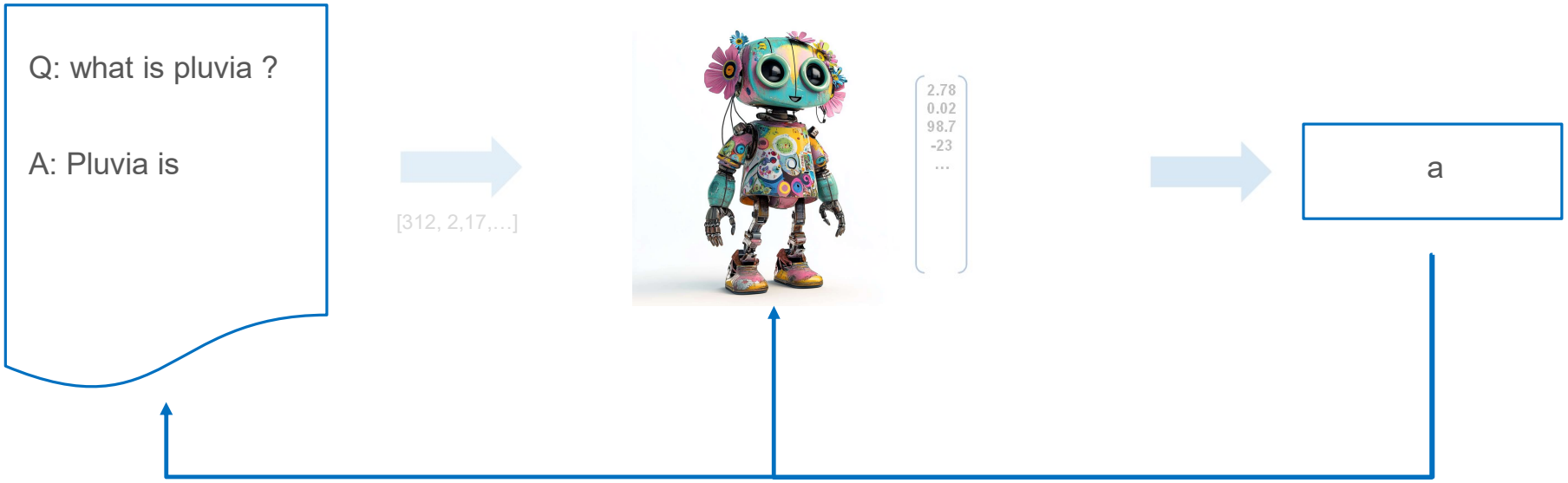


- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- **«the context»**

- 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 ?

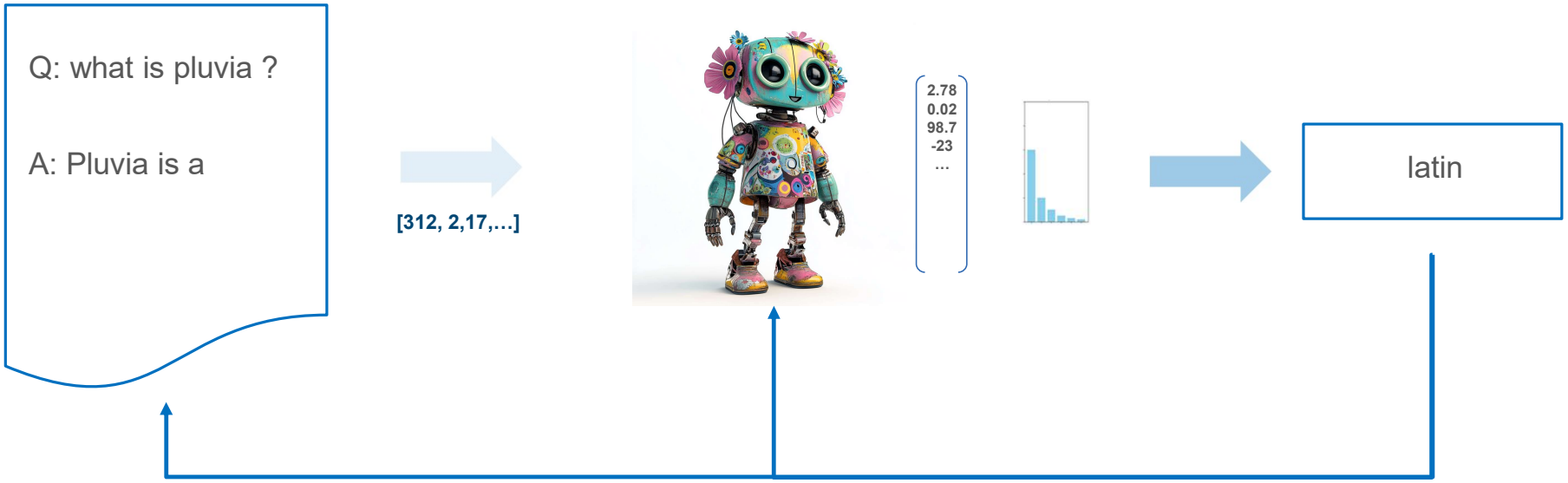


- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- «the context»

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

- Single «word»
- Depends on context and model
- «the token»
- **Represented as an int id**

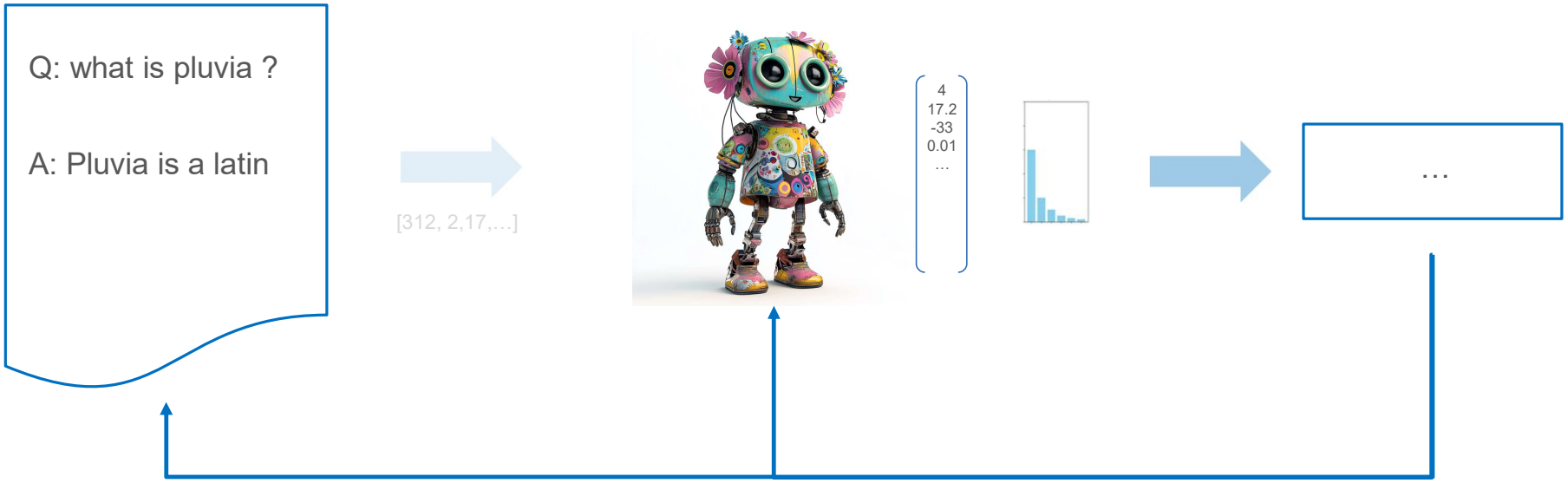
How does a Decoder Model work ?



- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- «the context»
- **A list of token ids**

- Trained on huge datasets
- Does not change
- Same for all users
- «the model»
- **Output a probability distribution over token ids**

- Single «word»
- Depends on context and model
- «the token»
- **Represented as an int id**

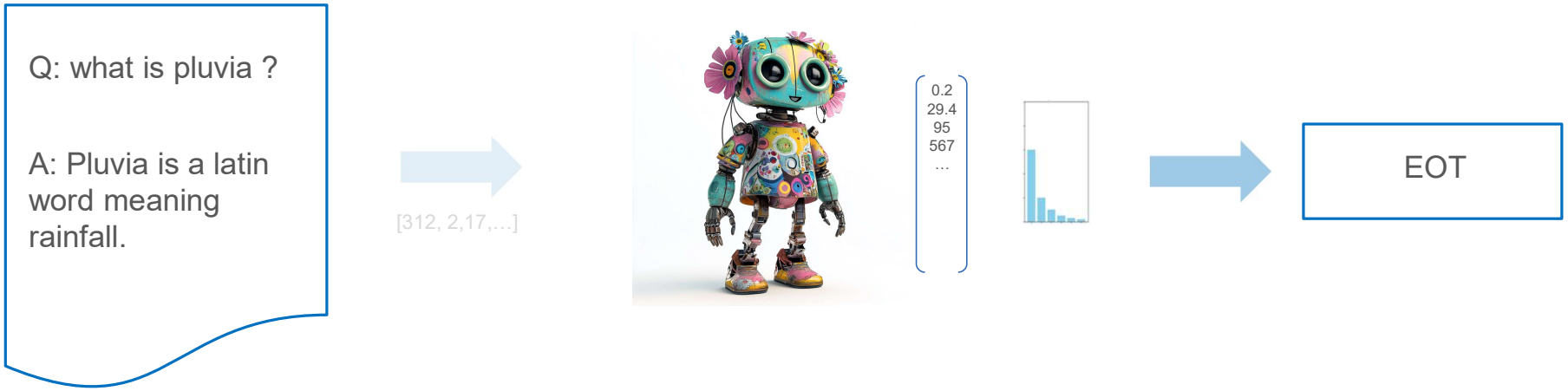


- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- «the context»
- **A list of token ids**

- Trained on huge datasets
- Does not change
- Same for all users
- «the model»
- **Output a probability distribution over token ids**

- Single «word»
- Depends on context and model
- «the token»
- **Represented as an int id**

How does a Decoder Model work ?



- Depends on users goal
- Unique for each chat & user
- Contains the chat history
- «the context»
- **A list of token ids**

- Trained on huge datasets
- Does not change
- Same for all users
- «the model»
- **Output a probability distribution over token ids**

- Single «word»
- Depends on context and model
- «the token»
- **Represented as an int id**

Decoder On-Prem Challenges

- Context sizes vary (depending on the Model)
  - with large contexts certain positions might be blind spots
- Memory consumption grows with context used
- Scaling to more than one parallel request

Inference on GPU only

Comparing suitable NVDIA GPUs

- T4/RTX 20 : https://en.wikipedia.org/wiki/Turing_(microarchitecture)
- V100 - professional variant of RTX 20 consumer line: https://en.wikipedia.org/wiki/Volta_(microarchitecture)
- A100/RTX 30 : https://en.wikipedia.org/wiki/Ampere_(microarchitecture)
- L4 /L40/RTX 40 : https://en.wikipedia.org/wiki/Ada_Lovelace_(microarchitecture)
- H100 - professional variant of RTX 40 consumer line: https://en.wikipedia.org/wiki/Hopper_(microarchitecture)

Commodity: available for small money

Architecture Decision: What can we work with?

Limiting factor is GPU RAM

- T4 : 16GB
- A100 : 40GB/80GB
- L4 : 24GB (L40: 48GB)

Limiting factor is GPU RAM: Quantization

- Typical resolution is 16 Bit
- Each parameter takes 2 Bytes on GPU memory
- Caching for context comes on top
- Varies with
  - length of context
  - architecture of model
  - batch size
- What about reducing resolution to 8 Bit or 4 Bit?
- Thus cutting memory requirement down to half or quarter?
- Overview: https://huggingface.co/docs/transformers/main/en/quantization/overview

Bitsandbytes:
Most straight forward approach to quantization

- https://huggingface.co/docs/text-generation-inference/conceptual/quantization#quantization-with-bitsandbytes
- Deep Dive: https://huggingface.co/blog/hf-bitsandbytes-integration
- Can go down to 4 Bits: https://huggingface.co/blog/4bit-transformers-bitsandbytes
- inference can be slower than more sophisticated methods (like GPTQ) or full FP16 precision
  - https://huggingface.co/blog/hf-bitsandbytes-integration#is-it-faster-than-native-models

Architecture Decision: Smaller Model vs Quantiztion

Key Driver multi language

Smaller Models

- Microsoft Phi 3.5 3.8B: https://huggingface.co/microsoft/Phi-3.5-mini-instruct
- Google Gemma 2 2.6B: https://huggingface.co/google/gemma-2-2b-it

Quantization

- https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct
- Quantized with bitsandbytes
- to 4-Bit: https://huggingface.co/docs/transformers/main/en/quantization/bitsandbytes?bnb=4-bit
- and 8-Bit: https://huggingface.co/docs/transformers/main/en/quantization/bitsandbytes?bnb=8-bit

Hands-On:
Quantize Meta-Llama 3.1 8B

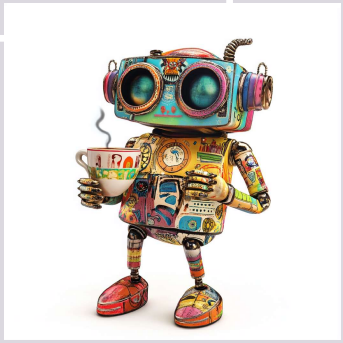
https://colab.research.google.com/github/DJCordhose/practical-llm/blob/main/Assessment.ipynb?hl=en

Local machine without NVIDIA GPU

- llama.cpp
  - https://github.com/ggerganov/llama.cpp/blob/master/README.md
  - https://www.theregister.com/2024/07/14/quantization_llm_feature/
  - Quantization and optimization
  - Optimized for Apple Silicon M1/M2/M3/M4
- Ollama
  - Simplifies usage of llama.cpp
  - https://ollama.com/
  - https://github.com/ollama/ollama
  - https://www.theregister.com/2024/03/17/ai_pc_local_llm/

60 Minuten

Coffee Break



Larger Decoder Models

Architecture Decision: Big Models on Heavy Hardware ?

- There are more powerful versions of OS decoder models available
  - Rival OpenAI GPT models
  - Support for major European languages
- Quantized versions will run on small GPUs, but far too slow for real world
  - Useful as demonstration only
- Those models will run on available hardware and dedicated inference server
  - **H100 GPUs are are expensive, but available**
  - Inference servers optimize for latency and throughput
    - https://huggingface.co/docs/text-generation-inference
    - https://developer.nvidia.com/nim
- We can get a preview
  - https://build.nvidia.com/explore/discover
  - https://huggingface.co/chat/

The future is already here, it is just not evenly distributed
- William Gibbson

Option: Mixtral 8x7B

- Good context length: 24K input, 8K output
- explicitly tuned for European languages (like French, Italian, German and Spanish)
- Mixture of experts
- only uses fraction of parameters at a time
- thus also bringing down KV-cache needs

Reference

- https://mistral.ai/news/mixtral-of-experts/
- https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
- Sparse Mixture of Experts (SMoE) Mixtral 8x7B: https://arxiv.org/abs/2401.04088

Option: Llama 3.1 70B

- Even better context length: 128k
- Supported languages: English, German, French, Italian, Portuguese, Hindi, Spanish, and Thai.
- Significantly better scores in European languages than 8B version
- Compared to Mixtral 8x7B
  - significantly better scores all over
  - Needs more memory and compute

Reference

- https://huggingface.co/meta-llama/Meta-Llama-3.1-70B-Instruct
- https://ai.meta.com/blog/meta-llama-3-1/
- https://llama.meta.com/

Big Model, Inference Server & GPU

- Model can be run as they are, just add a REST API
- Better: Inference Server (e.g. Huggingface text generation interface/TGI)
  - Batching of requests
  - Automatic usage of existing hardware
  - Optimization
  - Distribute model over multiple GPUs (Tensor parallelism)

https://huggingface.co/docs/text-generation-inference

It works:
Mixtral 8x7B on 2xH100 NVL using TGI

GB200 - Future successor to both Hopper &Ada Lovelace

- https://en.wikipedia.org/wiki/Blackwell_(microarchitecture)
- 2,5x faster than H100
- 2x memory
- Native support for 4 Bit resolution
- Sped up NVLink

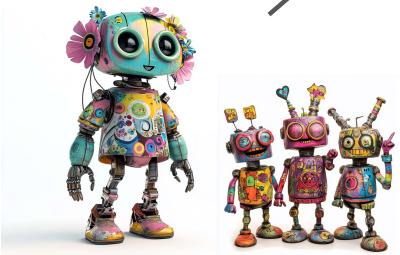
https://www.heise.de/news/Nvidias-neue-KI-Chips-Blackwell-GB200-und-schnelles-NVLink-9658475.html

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

Ground Truth

Human Eval

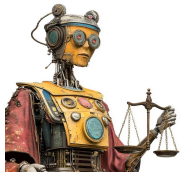
Statistics

Evaluation Criteria:

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

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

LLM as a judge: Idea

Generation

- Llm Input : “Why do you dislike writing texts ?”
- Llm Output : «Witing texts is painful, caus im making mitakes.”

Context

You are an expert on english language. Grade a students text with scores between 0 and 10.
Answer with a Json containing a score and a reason.

Student Text: Witing texts is painful, caus im making mitakes.

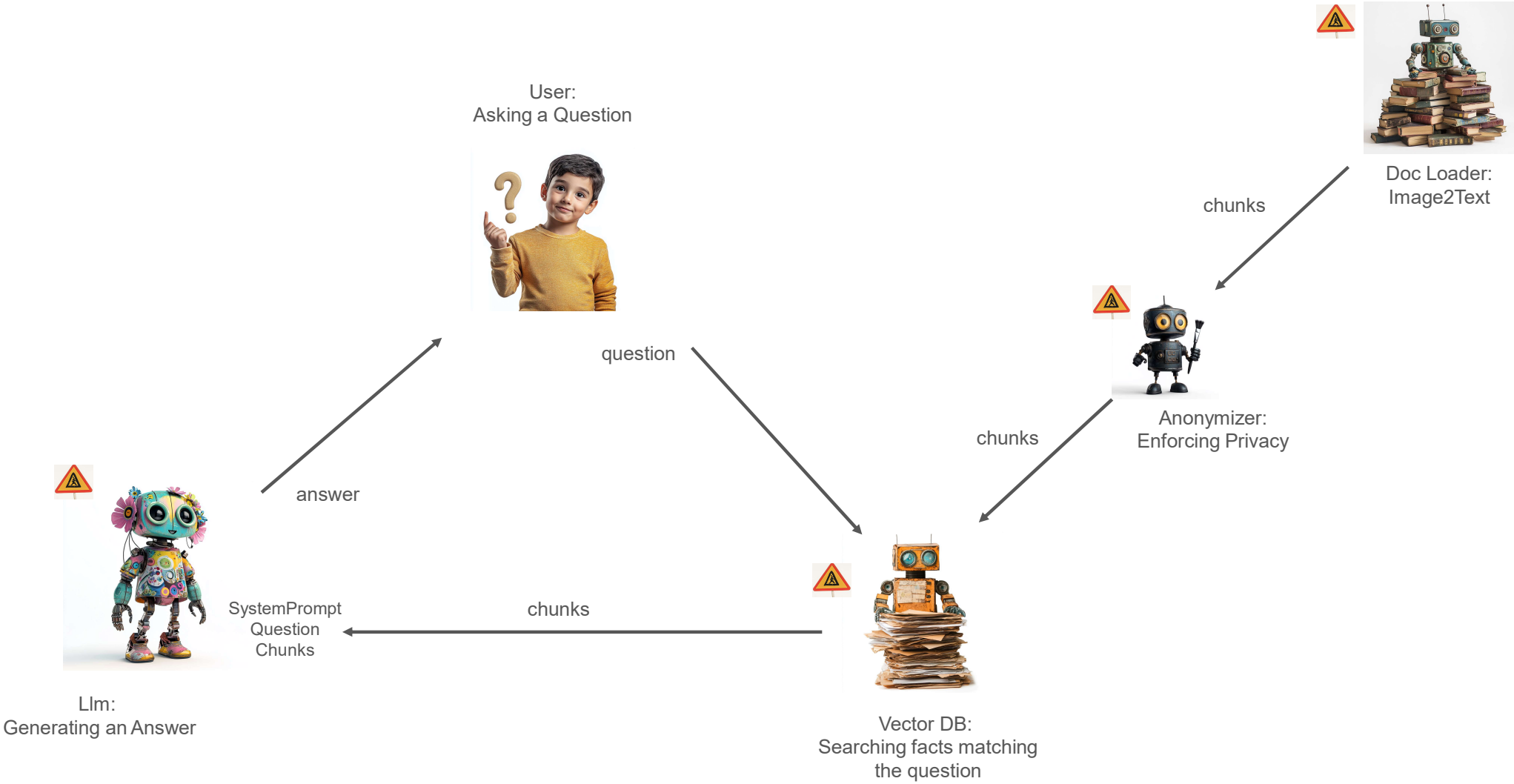
Json:

Result

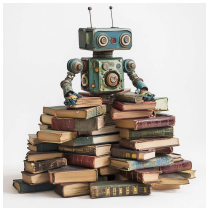
```
{
    'score’:    2,
    'reason’:   "The text contains multiple spelling errors, such as 'Witing' instead of 'Writing.“
}
```

Demo:
Evaluation on Prem Notebook

System Architecture



System Architecture: Evaluation



Doc Loader:
Image2Text

chunks

User:
Asking a Question

Anonymizer:
Enforcing Privacy

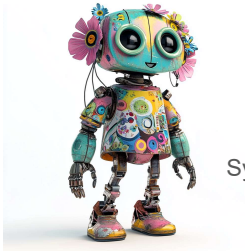
chunks

question

Contextual Relevance

Answer Relevance
Conciseness

answer



Llm:
Generating an Anwer

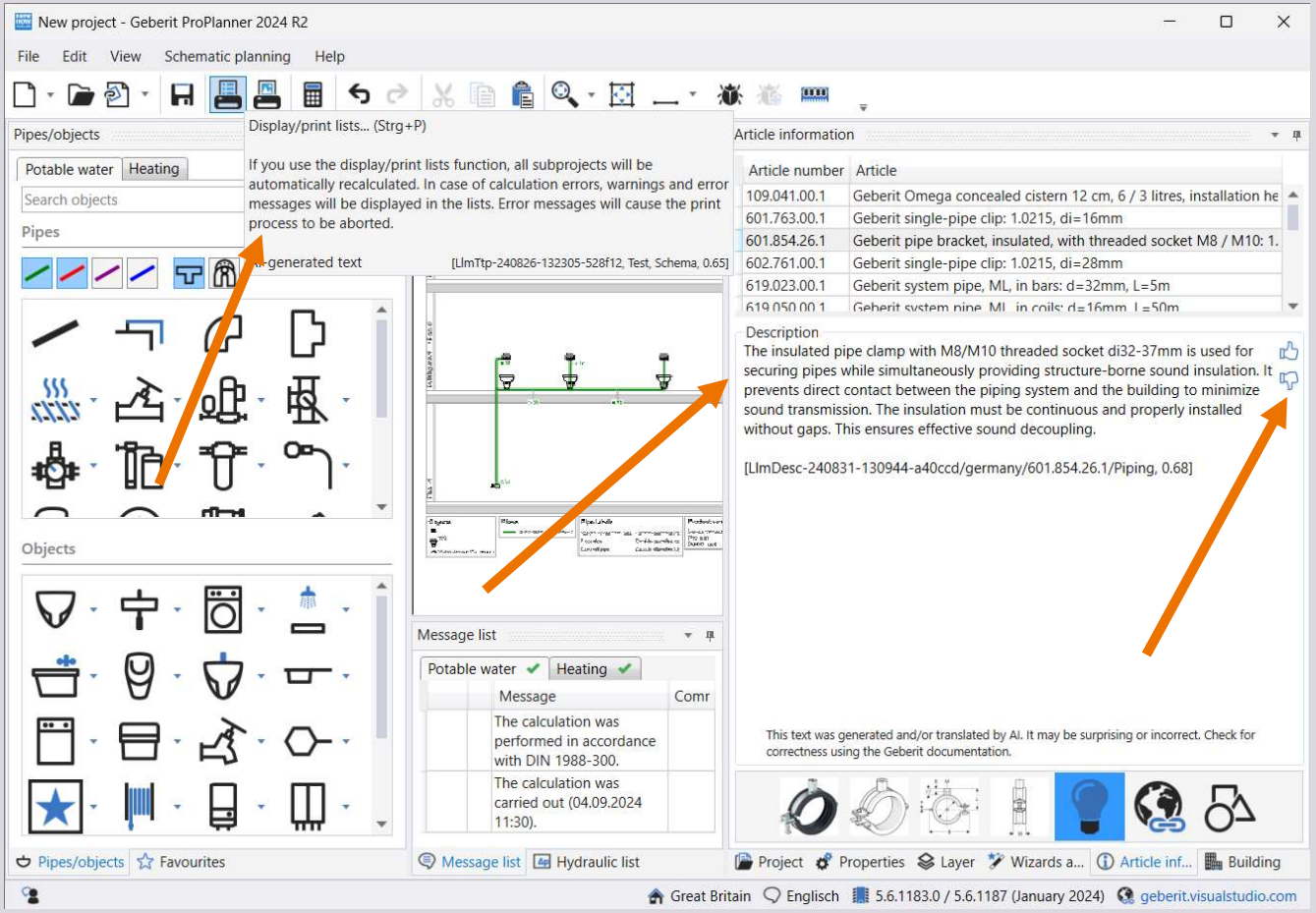
SystemPrompt
Question
Chunks

chunks

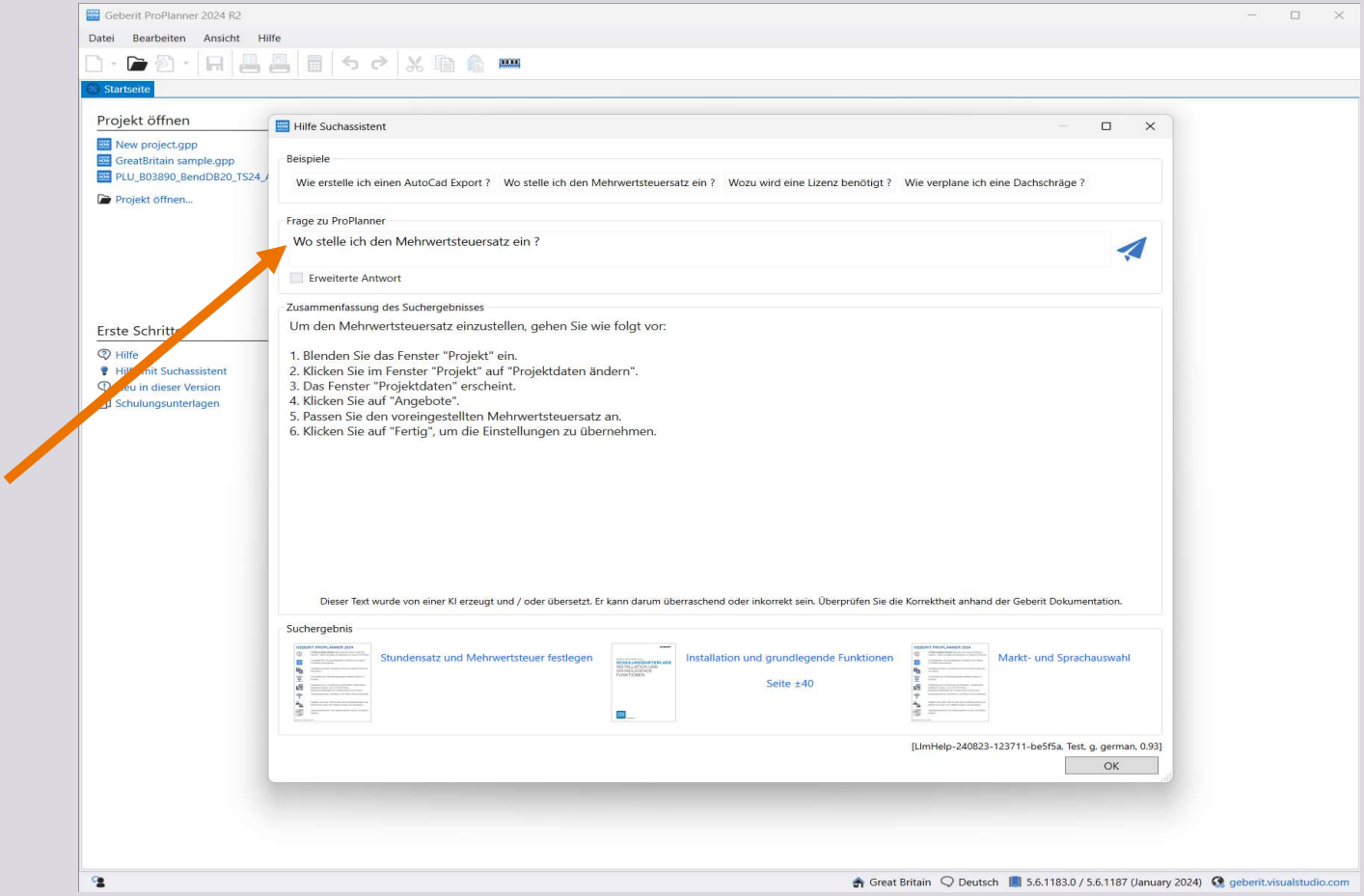
Vector DB:
Searching facts matching
the question

Faithfulness

Online Eval: Example



Online Eval: Example



Online Eval: Example


```
{
  "DtoTypeId": "Description",
  "DtoCreatedDate": "2024-08-31T13:09:44.4579038Z",
  "GeberitArtNo": "601.854.26.1",
  "RequestId": "LlmDesc-240831-130944-a40ccd",
  "SourceText": "Rohrschelle gedämmt mit Gewindemuffe M8/M10 di32-37mm",
  "TargetLanguage": "english",
  "TargetMarket": "germany",
  "TargetMaxSentences": 4,
  "TargetText": "The insulated pipe clamp with M8/M10 threaded socket di32-37mm is used for securing pipes while simultaneously providing structure-borne sound
  "Topic": "Piping",
  "RagSources": [
  "Context": {
  "RagEval": {
    "MetaData": {
      "Answer": "Die gedämmte Rohrschelle mit Gewindemuffe M8/M10 di32-37mm dient zur Befestigung von Rohren, während sie gleichzeitig eine Körperschalldämmung
      "CreatedDate": "2024-08-2[illegible]7:59:20.983207Z",
      "DeepEval": {
        "Answer_Relevancy": {
          "reason": "The score is 1.00 because the response directly addresses the purpose of the article without any irrelevant statements.",
          "score": 1.0
        },
        "Conciseness_(GEval)": {
          "reason": "The output is clear and relevant, but it could be more concise by reducing some repetitive phrases about sound insulation.",
          "score": 0.7
        },
        "Contextual_Relevancy": {
          "reason": "The score is 0.00 because the context discusses the Geberit Silent-db20 and unrelated technical specifications, failing to address the spec
          "score": 0.0
        },
        "Faithfulness": {
          "reason": "The score is 1.00 because there are no contradictions, indicating that the actual output perfectly aligns with the retrieval context.",
          "score": 1.0
        }
      },
      "ElapsedSeconds": 14.57,
      "EvalType": "deep_eval",
      "EvalVersion": "240811",
      "Input": "Wozu dient der Artikel Rohrschelle gedämmt mit Gewindemuffe M8/M10 di32-37mm ?"
    },
```

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

Wrap Up

Key takeaways

Collection of notebooks used

- Quantization:
  https://colab.research.google.com/github/DJCordhose/practical-llm/blob/main/Assessment.ipynb
- Evaluation:
  https://colab.research.google.com/github/DJCordhose/practical-llm/blob/main/Eval4pptx.ipynb

Thank you

Collection of notebooks used

- SetFit: https://colab.research.google.com/github/DJCordhose/practical-llm/blob/main/Assessment_SetFit.ipynb
- Microsoft Phi 3 3.8B: https://colab.research.google.com/github/DJCordhose/practical-llm/blob/main/Assessment_Phi_3_mini_T4.ipynb
- Google Gemma 2 2B: https://colab.research.google.com/github/DJCordhose/practical-llm/blob/main/Assessment_Gemma_2_2B_T4.ipynb
- Meta Llama 3.1 8B - Quantized to 4-Bit and 8-Bit: https://colab.research.google.com/github/DJCordhose/practical-llm/blob/main/Assessment_Llama_3.1_8B_Quantize_T4.ipynb
- Meta Llama 3.1 8B: https://colab.research.google.com/github/DJCordhose/practical-llm/blob/main/Assessment_Llama_3.1_8B_Full_T4.ipynb
- Mixtral 8x7B with extreme quantization: https://github.com/DJCordhose/practical-llm/blob/main/Assessment_Mixtral_8x7B.ipynb