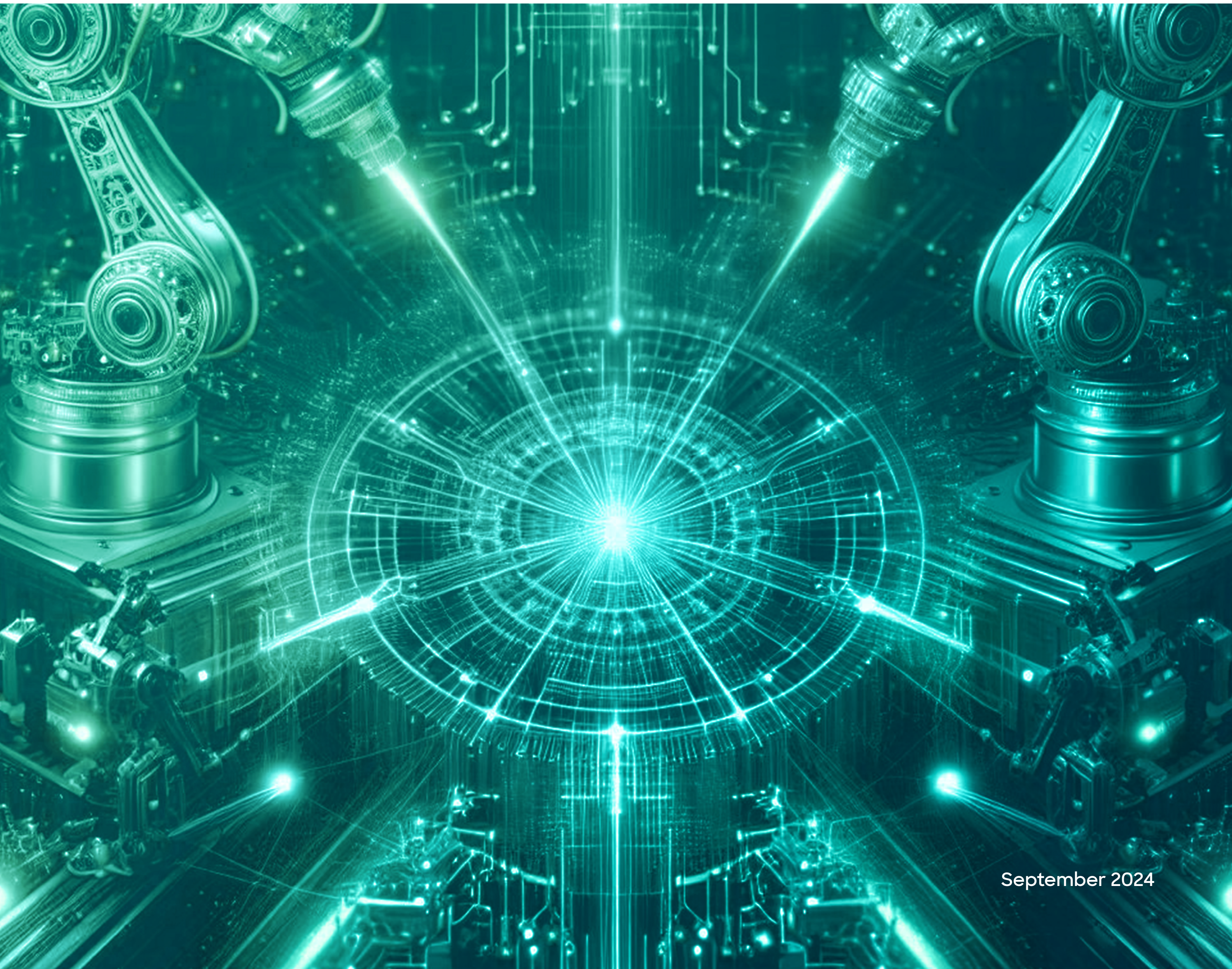
deepset

ROHDE&SCHWARZ
Make ideas real

initiative for applied artificial intelligence

How Do I Optimize the Dynamic Range of an FSW Signal and Spectrum Analyzer?

A RAG Use Case Study in Wireless Test & Measurement: Retrieval Fine-Tuning and Tables as Images



September 2024

Contents

Foreword 4

Executive Summary 5

Success Begins With Strong Collaboration 7

Chapter 1: Introduction and Overview 9

From Where We Started: Initial Implementation by Rohde & Schwarz 10

Understanding the Problem: Challenges We Have Addressed 12

New Features to Open Up New Possibilities: What We Have Achieved - And How 14

Chapter 2: Diving Deep 16

Understanding Tabular Data Through Vision: Interpreting Tables as Images 17

Enhancing Neural Understanding of Specialized Data: Retriever Fine-Tuning 19

From Manual to Automatic Evaluation: Finding Reliable Automatic Metrics 23

Chapter 3: Learnings and Future Perspectives 25

Insights and Future Perspectives 26

References 28
Authors 29
Contributors 30
About appliedAI Initiative GmbH 31
About deepset 32
About Rohde & Schwarz 33
Acknowledgement 34

Foreword

As organizations across industries increasingly adopt and rely on Generative AI, the demand for trustworthy solutions that allow organization-wide scaling and value creation has never been more critical. At the forefront of this movement is *Trustworthy Retrieval-Augmented Generation (RAG)*, a powerful approach that transforms business processes and customer experiences at scale with top-tier accuracy, fast-time-to-value, and future-proof adaptability.

This case study paper reflects the joint vision of deepset and appliedAI to drive the industrialization of Trustworthy RAG and GenAI solutions together, combining our respective strengths:

1. **Deepset's cutting-edge enterprise platform technology,** based on the widely adopted Haystack framework, to customize generative AI solutions for production-ready business applications, and

2. **appliedAI's deep expertise in professional services,** to advance the AI journey of companies focusing AI Transformation and AI implementation in a structured, scalable, and sustainable way.

Together, our focus is on ensuring controllability, explainability, and transparency in AI systems—key differentiators that allow businesses to trust their AI solutions. For enterprises facing stringent data security, confidentiality, and regulatory requirements, this collaboration ensures scalable and secure Trustworthy RAG implementations for the most critical business processes and services.

We invite you to explore the power of *Trustworthy RAG* through the following case study, where real-world examples demonstrate how these solutions are transforming businesses and addressing complex challenges.

Dr. Andreas Liebl
CEO, appliedAI Initiative

Milos Rusic
Co-founder & CEO, deepset

Executive Summary

Use Case Background

- Rohde & Schwarz manufactures a wide range of high-precision devices, including oscilloscopes, spectrum analyzers, and communication testers
- The devices and corresponding specifications are backed up by a **rich resource of domain-specific data, ranging from user manuals of more than 1000 pages for a single device to data sheets containing complex tabular data structures**
- Manually referencing this comprehensive data for troubleshooting, calibrations, and device setup in specialized scenarios is often **labor-intensive and time-consuming**. Consequently, optimizing knowledge processing and management holds significant potential for enhancing efficiency and reliability
- **Retrieval-augmented generation (RAG)**, combining the natural language processing power of large language models (LLMs) with company-specific information through information retrieval, provides a **viable solution for making data easily accessible in a trustworthy way**

Challenges and Solutions

Complex Tabular Data Interpretation

Challenge
- Devices can have a large number of specifications that are documented in **complex tabular structures**
- Table parsing is a complex process where naive parsing can lead to loss of tabular structure and in turn confusion during answer generation

Solution
- We developed a **vision pipeline** where tables are interpreted as images and fed to a vision language model
- The vision pipeline allows us to interpret the table as it was meant to be and is activated if an answer can not be correctly generated based on text input

Domain-Specific Queries and Data

Challenge
- Questions and information are formulated in a **domain-specific language and terminology**
- The resulting shift between training and testing distribution of retrieval models can lead to sub-optimal retrieval and, consequently, insufficient answer correctness

Solution
- We **fine-tuned a neural retrieval model** utilizing LLMs for synthetic query generation and data labeling
- Retriever fine-tuning based on synthetic data provides a robust solution to enhance retrieval performance in specific domains

Ambiguous or Underspecified Queries

Challenge
- Queries can be based on **prior knowledge or require context to understand**, such as asking for a specific feature that might exist for multiple devices
- The uncertainty in the query hinders the ability to deliver a highly confident and accurate response directly

Solution
- We developed a **chat feature** that rewrites queries based on the chat history, ensuring they are independent of prior interactions.
- Chat allows the model to resolve uncertainties in the query through dialogue interactions, enhancing the correctness and user experience

Multilingual Queries and Data

Challenge
- Globally operating companies encounter **user queries and available data sources in multiple languages**
- Focusing only on a single language limits the accessibility of users and documents to retrieve

Solution
- We integrated **multilingual retrieval and reranker models** into our RAG pipeline to support multilingual capabilities
- Multilinguality enables the understanding of queries in various languages and the retrieval of documents that may be in a different language than the query

Executive Summary

Results

- **Overall improvement:** our optimized RAG solution **improved the number of correctly answered questions by 46%** (manual evaluation of 79 questions in total)
- The overall improvement can be attributed to multiple factors, such as **changing the LLM for answer synthesis, changing and fine-tuning the retrieval model, and adding a reranker and vision pipeline** into the RAG system
- Besides pipeline improvements, we **boosted the scalability of data generation and system evaluation** through a mixture of minimal manual effort and utilization of the generative power of LLMs

Recommendations

Build on a modular RAG framework technology that

- seamlessly integrates of specialized components to meet specific requirements
- ensures transparency in data processing throughout the RAG pipeline, providing insights into how results are generated not be correctly generated based on text input

Put emphasis on core RAG components and consider

- retrieval fine-tuning on synthetic data as a robust solution for improving domain-specific retrieval performance
- different LLM choices to optimize cost, efficiency, and output behavior

Scale-up evaluation through a hybrid, semi-automated framework

- consisting of minimal expert effort and LLM automation
- to assess the advancements within fast iteration cycles

Combine the strength of in-domain and RAG experts to

- carefully assess incorrectly answered questions and derive weak points in the RAG system that can be rapidly advanced upon
- enhance the user experience by exploring specific user needs and define a final product appearance and operation

Success Begins With Strong Collaboration

Powering Success Through Expert Collaboration

Why Combining In-Domain and RAG Experts Matters

- **Multiple Expertise:** In-domain experts understand the specific challenges of the domain, while RAG experts bring technical skills to build and optimize RAG pipelines
- **Efficient Problem Solving:** The collaboration ensures that both the problem and the technical solution are fully understood, leading to more effective and tailored solutions
- **Faster Development:** RAG experts can rapidly prototype while domain experts guide the refinement, resulting in accelerated progress
- **Cross-Project Insights:** RAG experts leverage experience from other projects, applying best practices and innovative solutions to new challenges
- **Improved Outcomes:** Combining deep domain knowledge with advanced technical capabilities enhances the quality and relevance of the final solution

The Working Method to Thrive

- **Rapid Iteration Cycles:** Focus on end-user needs and continuous feedback, ensuring the team can quickly adapt and refine solutions to meet user requirements
- **Informed Prioritization:** Features are prioritized based on value and complexity, maximizing development efficiency and focusing on high-impact improvements
- **Workshops for In-Depth Insights:** Expert users test RAG pipelines in realistic settings, providing valuable insights that help steer development in the right direction
- **Ongoing Evaluation:** Continuous assessment through expert reviews and automated metrics ensures the solution evolves with measurable improvements

Success Begins With Strong Collaboration

Note on appliedAI's behalf

Collaborative Use Case Study for appliedAI Partners to Address Challenges in the Implementation of AI Use Cases Together

What appliedAI provides:

- In-depth AI expertise encompassing engineering, the field's latest research and best practice solutions
- Project strategy and coordination
- Access to and involvement in the appliedAI ecosystem for additional support, including methods, technology, and knowledge

Joint Case Study

What an appliedAI partner brings:

- Use case, data, and optionally, an existing prototype for benchmarking
- Domain expertise, with optional AI expertise
- Access to internal subject matter experts for additional support, including domain knowledge, user experience, and evaluation

Value for the partner:

- **Benchmarking and advancing** internal proposals and prototypes in collaboration with appliedAI to strengthen **decision-making for implementation**
- Receiving concrete **recommendations, best practices and prototype implementations** to ensure a successful MVP implementation
- Promoting results together within the ecosystem to gain **visibility as a leader in AI**

Chapter 1: Introduction and Overview

From Where We Started: *Initial Implementation by Rohde & Schwarz*

Use Case Background and Motivation

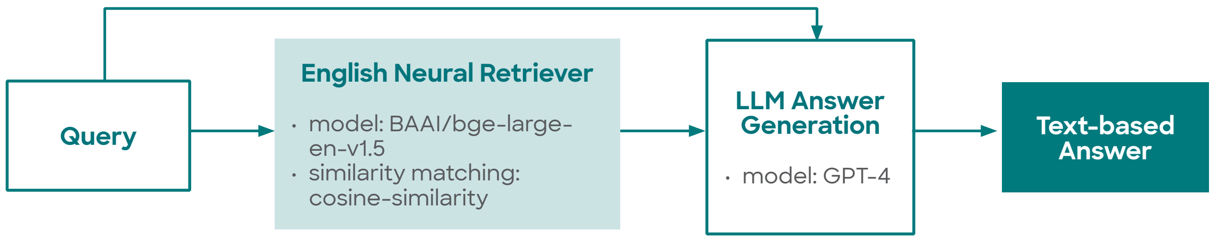
- Rohde & Schwarz manufactures a wide range of high-precision devices, including oscilloscopes, spectrum analyzers, and communication testers
- The devices and corresponding specifications are backed up by a **rich resource of domain-specific data, ranging from user manuals of more than 1000 pages for a single device to data sheets containing complex tabular data structures**
- Manually referencing this comprehensive data for troubleshooting, calibrations, and device setup in specialized scenarios is often **labor-intensive and time-consuming.** Consequently, optimizing knowledge processing and management holds significant potential for enhancing efficiency and reliability
- **Retrieval-augmented generation (RAG)**, combining the natural language processing power of large language models (LLMs) with company-specific information through information retrieval provides a **viable solution for making data easily accessible in a trustworthy way**

Initial Implementation by Rohde & Schwarz

Motivated by the opportunities of RAG, at the end of 2023, the ML experts at Rohde & Schwarz developed and deployed an early RAG prototype based on a simplified architecture. This enabled early adoption for multiple internal use-cases. Although the architecture of the internal deployment is constantly evolving and improving, we take this particular deployment as a baseline for performance evaluation for this use-case study. The prototype includes:

- **English neural retriever** for retrieving the most similar documents to a given query
- **LLM answer generator** for answering the query based on the retrieved documents

To evaluate the prototype, a Q&A dataset comprising 79 questions and corresponding correct answers was created. A manual evaluation by domain experts of Rohde & Schwarz resulted in **34 out of 79 correctly answered questions.**



From Where We Started:
Initial Implementation by Rohde & Schwarz

Info: A Brief Introduction to RAG

The Challenges of LLMs

LLMs open up new possibilities due to their natural language understanding, generation, and reasoning capabilities. Nevertheless, they come with certain challenges:

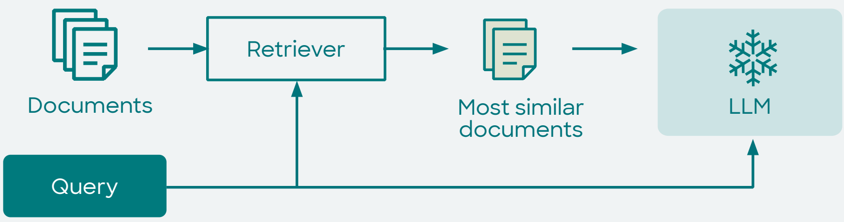
- **Knowledge cutoff:** LLM knowledge is limited to the (potentially outdated) training data
- **Incomplete knowledge:** LLMs have no knowledge of private, proprietary data
- **Lack of trustworthiness:** LLM answers do not provide a grounded source

LLMs cannot answer company-specific questions out of the box in a grounded way. Instead, LLMs need to be provided with the **right, context-specific company information**.

RAG: Connecting LLMs and Company Information

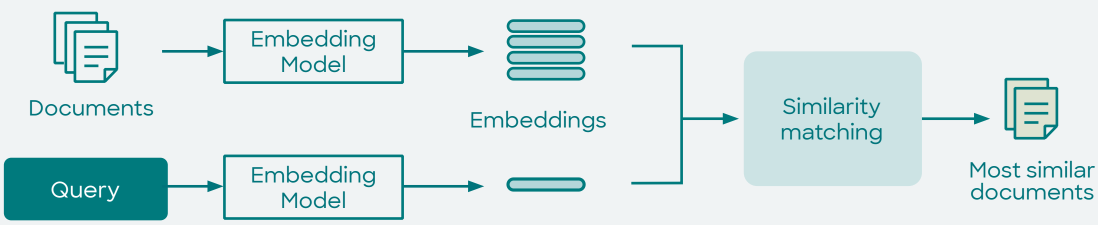
Retrieval-augmented generation [1, 2] is a paradigm that augments the input of an LLM with grounded, context-specific (company) information retrieved by a retrieval model, thereby addressing the challenges of solely utilizing LLMs.

A simple RAG pipeline



- Based on a query, RAG applications first find documents within the corpus that are relevant to the query
- The relevant documents, together with the query, are fed to the LLM to generate an answer grounded in company-specific information

Neural Retrieval Models



- In contrast to keyword-based retrievers that compare query and documents lexically, neural retrievers compare query and documents semantically utilizing embeddings. **Neural retrievers** find relevant documents to a query in two steps:
  - query and documents are transformed into a **numerical vector representation** (embedding)
  - query embedding and document embeddings are compared, returning the documents with the **highest similarity to the query** (similarity matching)
- Embeddings are constructed through an embedding model (neural network)

Understanding the Problem: *Challenges We Have Addressed*

After analyzing the data and use cases, we formulated the following four key challenges within our use case study. We emphasize that while these challenges were derived with respect to data from Rohde & Schwarz, they are representative of many specialized domain applications.

We refer to the following resources, accessible via the embedded links, to get a tangible experience of the encountered data:

- Tables: Rohde & Schwarz FSVA3000 Specifications
- User manual: Rohde & Schwarz FSVA3000 User Manual

Tabular Data

Rohde & Schwarz devices can have a large amount of specifications that are documented in **complex tabular structures.** This tabular data can be complex and difficult to parse into pure textual representations. Naive parsing often leads to loss of tabular structure, and in turn, possible confusion during answer generation.

Absolute level uncertainty at 64 MHz	RBW = 10 kHz, level –10 dBm, reference level –10 dBm, RF attenuation = 10 dB	
	+20 °C to +30 °C	< 0.2 dB (σ = 0.07 dB)
	0 °C to +50 °C	< 0.35 dB (σ = 0.12 dB)
Frequency response referenced to 64 MHz	RF attenuation = 10 dB, 20 dB, 30 dB, 40 dB, RF preamplifier = off, electronic attenuator off, +20 °C to +30 °C	
	9 kHz ≤ f < 10 MHz	< 0.5 dB (σ = 0.17 dB)
	10 MHz ≤ f < 3.6 GHz	< 0.3 dB (σ = 0.10 dB)
	3.6 GHz ≤ f ≤ 7.5 GHz	< 0.5 dB (σ = 0.17 dB)
	7.5 GHz < f ≤ 13.6 GHz, span < 1 GHz	< 1.5 dB (σ = 0.5 dB)
	13.6 GHz < f ≤ 30 GHz, span < 1 GHz	< 2.0 dB (σ = 0.66 dB)
	30 GHz < f ≤ 43.5 GHz, span < 1 GHz	< 2.5 dB (σ = 0.83 dB)
	43.5 GHz < f ≤ 50 GHz, span < 1 GHz	< 3.5 dB (σ = 1.16 dB)
	any setting of RF attenuation, RF preamplifier = off, 0 °C to +50 °C	
	9 kHz ≤ f < 3.6 GHz	< 1.0 dB (σ = 0.33 dB)
	3.6 GHz ≤ f ≤ 7.5 GHz	< 1.5 dB (σ = 0.5 dB)
	7.5 GHz < f ≤ 13.6 GHz	< 2.5 dB (σ = 0.83 dB)

Pre-retrieval + Retrieval Stage

Domain-Specific Queries and Data

Questions and information are formulated in a **domain-specific language and terminology** that is very different from what was observed in the training data of a neural retrieval model, possibly leading to sub-optimal retrieval results.

What is the first intermediate frequency (IF) on the Rohde & Schwarz FPL1000?

Parameters:
<State> ON | OFF | 0 | 1
*RST: 1

Example: `INP:ATT:AUTO ON`
Couples the attenuation to the reference level.

Manual operation: See "Attenuation Mode / Value" on page 59

INPut:EATT <Attenuation>

Defines an electronic attenuation manually. Automatic mode must be switched off (`INP:EATT:AUTO OFF`, see `INPut:EATT:AUTO` on page 148).

Retrieval Stage

Ambiguous or Underspecified Queries

Queries can be based on **prior knowledge or require context to understand,** such as asking for a specific feature that might exist for different device types.

I would like to test the performance of my antenna. Which Rohde & Schwarz product do you recommend?

User Query Stage

Multilingual Queries and Data

As a globally operating company, Rohde & Schwarz addresses an **international customer base who speak a multitude of different languages.** In addition, the vast amount of document data exists in various, but not always all, languages.

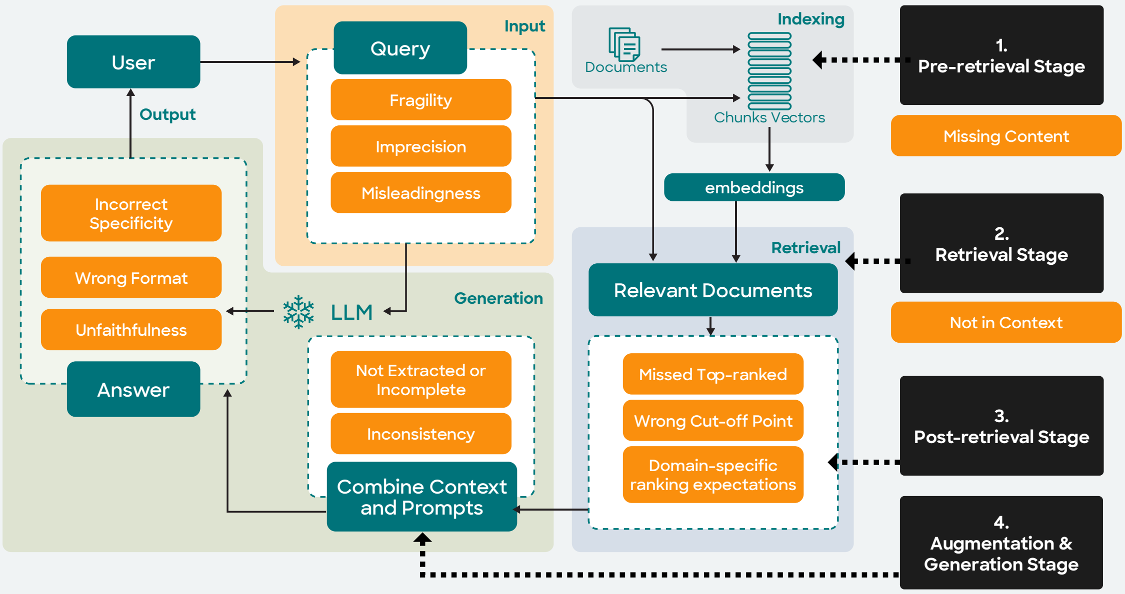
- Was ist der Unterschied zwischen den SCPI Befehlen *OPC? und *WAI
- How can the noise figure of the ZNA Rx path be decreased? Which options are required to do so?

Retrieval Stage

Understanding the Problem: *Challenges We Have Addressed*

Info: General Challenges in a RAG System [1, 3]

- **User queries** can be imprecise, misleading, and assume context information or prior knowledge
- **Pre-retrieval** includes document parsing and indexing, which can suffer from missing content and the need to parse complex structures within documents, such as tables and images
- **Retrieval** is a critical part that needs to ensure that the right context is retrieved
- **Post-retrieval,** typically including reranking, can suffer from wrong cut-off points and missing domain-specific ranking expectations
- **Augmentation and generation** can suffer from inconsistencies in retrieved document contents and unfaithful or incorrect answers (even for the correctly retrieved context)



cf. [1, 3]

New Features to Open Up New Possibilities: *What We Have Achieved - And How*

Visual Pipeline for Tabular Understanding

In order to omit the complex parsing process necessary to preserve tabular structure, we interpret tables as images and utilize the capabilities of multi-modal LLMs. To this end, we developed a **vision pipeline** to enhance tabular processing beyond sole textual understanding, observing the tables as they were meant to be. The vision pipeline acts as a fallback option in case no correct text-based answer is generated..

Improving Retrieval Through Fine-Tuning

In order to enhance understanding of domain-specific queries and data during retrieval, we perform **fine-tuning of an embedding model**. To this end, we created synthetic queries and automatically labeled test sets leveraging the generative power of LLMs.

Chat Functionality for Dialogue Interactions

In order to resolve uncertainty that stems from ambiguous or underspecified queries, we unlocked **chat functionality** by incorporating the chat history into the current query with query rewriting. Chat allows the system to clarify what the user meant in an interactive dialogue. Moreover, chat can be used to answer more complex queries that go beyond a single turn and for more extensive testing of the system pipeline.

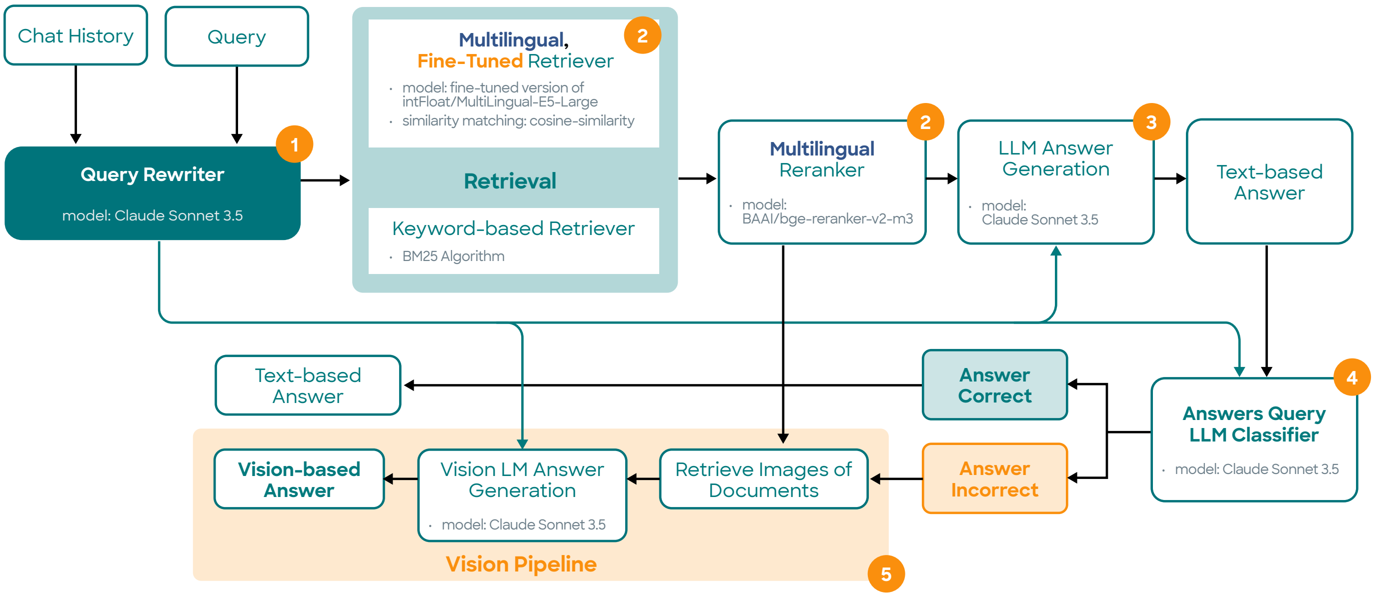
Multilingual Retrieval and Reranking

In order to support languages and data sources in multiple languages, we utilized **multilingual retriever and reranker models** within our RAG pipeline. More specifically, we used the multilingual IntFloat/E5-Large model for retrieval and BAAI/bge-reranker-v2-m3 for reranking.

"The AI shouldn't just answer; it should do research first to determine which of the answers are the best."

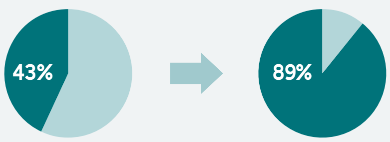
Jenson Huang
CEO, NVIDIA

New Features to Open Up New Possibilities: *What We Have Achieved – And How*



1. The query gets rewritten utilizing the chat history. This is an important step to enable chat-based interactions because the query might depend on context in the chat history, which would otherwise complicate the retrieval process
2. We leverage a multilingual, fine-tuned retriever model as well as a multilingual reranker for document retrieval and reranking (details on retriever fine-tuning can be found in section *Enhancing Neural Understanding of Specialized Data: Retriever Fine-Tuning*)
3. A text-based LLM produces an initial answer based on the retrieved documents and the query
4. The text-based answer will be assessed for correctness using an LLM classifier. In case the answer is deemed incorrect, a vision pipeline will be triggered
5. The vision pipeline obtains images based on the reranked documents and generates an answer (details can be found in section *Understanding Tabular Data Through Vision: Interpreting Tables as Images*)

Answer Correctness Improvement



34/79 correctly answered questions

70/79 correctly answered questions

Number of Indexed Documents

~450,000
documents were indexed

Number of Generated Queries

~150,000
synthetic queries with a cost of ~20$

Chapter 2: Diving Deep

Understanding Tabular Data Through Vision: *Interpreting Tables as Images*

Motivation

Important information is oftentimes stored in **large tabular data,** encompassing **complex structures** with multi-columns and various content types. Accurately parsing tabular information into textual data is complicated, and naive parsing often leads to the loss of the tabular structure, which is an important element in table comprehension. While text-only LLMs require textual input directly, humans use their vision capabilities to understand tabular structure. The rise of **multi-modal LLMs** opens up new possibilities for table understanding that omits the complex parsing process and interprets tables as they were meant to be.

Absolute level uncertainty at 64 MHz	RBW = 10 kHz, level –10 dBm, reference level –10 dBm, RF attenuation = 10 dB	
	+20 °C to +30 °C	< 0.2 dB (σ = 0.07 dB)
	0 °C to +50 °C	< 0.35 dB (σ = 0.12 dB)
Frequency response referenced to 64 MHz	RF attenuation = 10 dB, 20 dB, 30 dB, 40 dB, RF preamplifier = off, electronic attenuator off, +20 °C to +30 °C	
	9 kHz ≤ f < 10 MHz	< 0.5 dB (σ = 0.17 dB)
	10 MHz ≤ f < 3.6 GHz	< 0.3 dB (σ = 0.10 dB)
	3.6 GHz ≤ f ≤ 7.5 GHz	< 0.5 dB (σ = 0.17 dB)
	7.5 GHz < f ≤ 13.6 GHz, span < 1 GHz	< 1.5 dB (σ = 0.5 dB)
	13.6 GHz < f ≤ 30 GHz, span < 1 GHz	< 2.0 dB (σ = 0.66 dB)
	30 GHz < f ≤ 43.5 GHz, span < 1 GHz	< 2.5 dB (σ = 0.83 dB)
	43.5 GHz < f ≤ 50 GHz, span < 1 GHz	< 3.5 dB (σ = 1.16 dB)
	any setting of RF attenuation, RF preamplifier = off, 0 °C to +50 °C	
	9 kHz ≤ f < 3.6 GHz	< 1.0 dB (σ = 0.33 dB)
	3.6 GHz ≤ f ≤ 7.5 GHz	< 1.5 dB (σ = 0.5 dB)
	7.5 GHz < f ≤ 13.6 GHz	< 2.5 dB (σ = 0.83 dB)

Tabular data extracted from a Rohde & Schwarz document.

Integrating Vision Into a RAG Pipeline

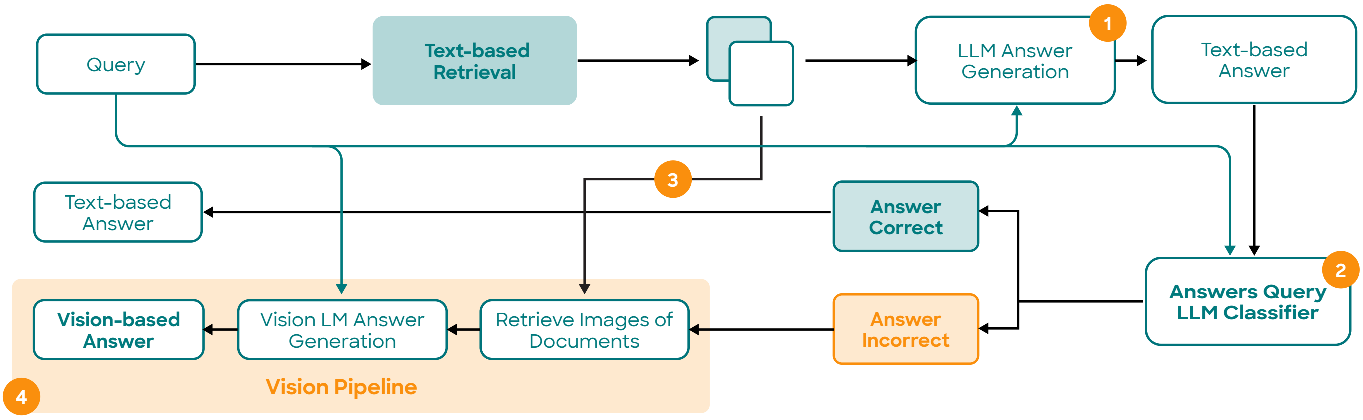
Our goal is to omit the complex parsing process of tabular data and instead utilize multi-modal LLMs for tabular understanding during query inference. To this end, we interpret tables as images and provide these images as input to a vision LLM. We integrate vision into a text-based RAG pipeline through a sequential process:

1. For a given query, we generate an answer using the **text-based RAG pipeline**
2. An **LLM classifier** obtains the generated answer and **assesses its correctness**
   - if the answer is deemed correct, it will be sent to the user
   - if the answer is deemed incorrect, it will be sent to the visual pipeline

If the vision pipeline is triggered:

3. For every document retrieved in the text-based pipeline, we extract the page it is located at as an image (this means image retrieval still requires text-based retrieval)
4. We provide the set of obtained images to a vision LLM for answer generation

Hence, the vision pipeline acts as a fallback option in cases where a text-based answer is insufficient. We emphasize here that while our goal is to omit the complex table parsing, the approach still requires text-based parsing since images are retrieved based on textual retrieval.



Understanding Tabular Data Through Vision: *Interpreting Tables as Images*

How Much Does Vision Help?

- **Adding a vision pipeline can enhance results considerably,** but its necessity depends on the strength of the text-based pipeline components
  - In an intermediate version of our text-based pipeline, adding the vision pipeline led to 13 more correctly answered questions
  - In our fully optimized pipeline, the vision pipeline was triggered only 3 out of 79 times due to enhanced retrieval and LLM components and could not deliver better answers due to the complexity of the questions. We thus recommend focusing first on enhancing the text-based pipeline until the additional requirement is identified
- **Vision output can lead to hallucinations on difficult questions or if the image quality is low**
  - Possible solution: provide the image and parsed content to the vision LM or zoomed-in screenshots to increase the image resolution of relevant tables
- **Based on our manual evaluation, visual QA:**
  - worked well when asking questions about specific values within table cells
  - struggled when needing to compare multiple values within a table or across tables
  - on its own led to inferior results compared to text-based pipelines, which is why we followed a sequential approach

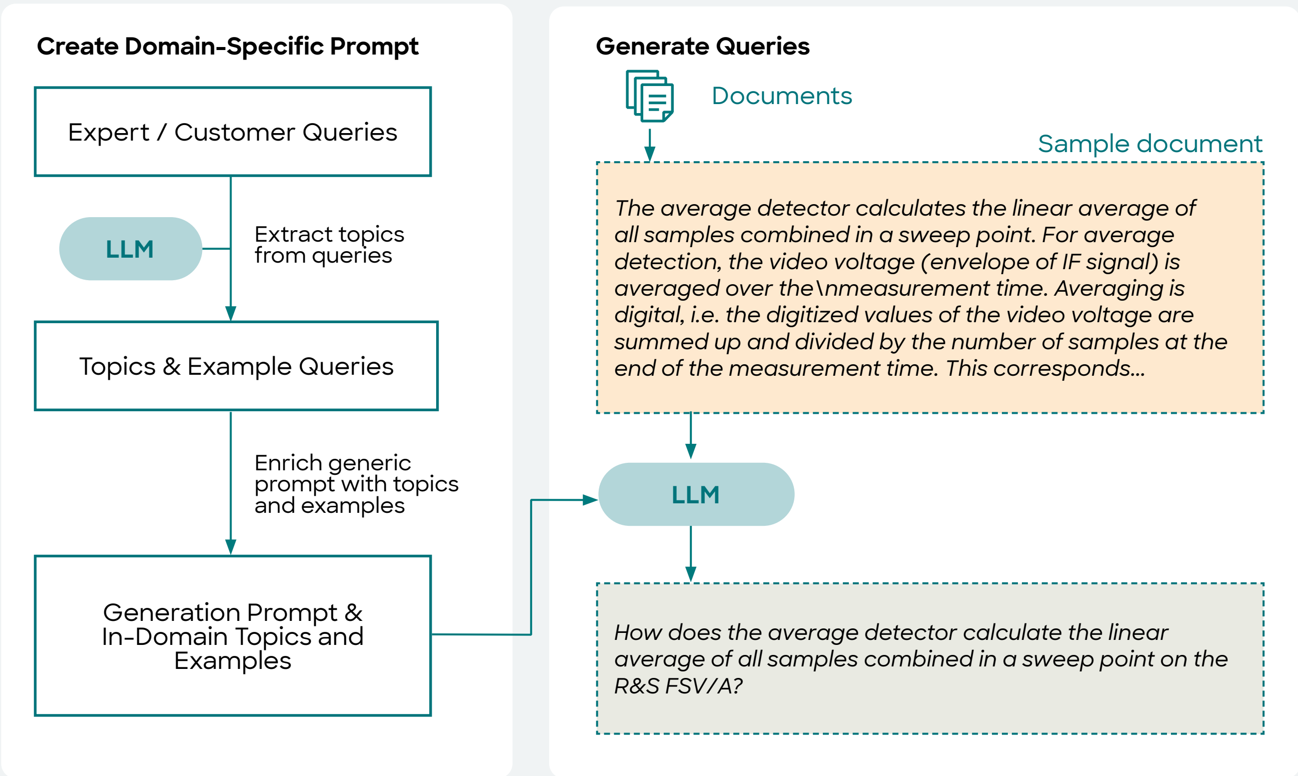
Enhancing Neural Understanding of Specialized Data: *Retriever Fine-Tuning*

Motivation

Embedding models are crucial for successful RAG applications, but they are often trained on general and public knowledge, limiting their effectiveness for company or domain-specific adoption. **Customizing embedding models through fine-tuning** can significantly boost the retrieval performance of domain-specific RAG applications, but acquiring domain-specific data for fine-tuning is difficult and cannot be done manually at scale. While domain-specific documents, typically the only source available at scale, are not enough for fine-tuning, the generative power and in-context learning capabilities of LLMs create new possibilities through **synthetic data generation and automatic labeling.** Our goal is to generate synthetic training data for embedding model fine-tuning as well as generating test sets for evaluating the retrieval performance.

Preparing for Fine-Tuning: Synthetic Queries, Hard Negative Documents, and Labeled Test sets

Creating Synthetic Queries



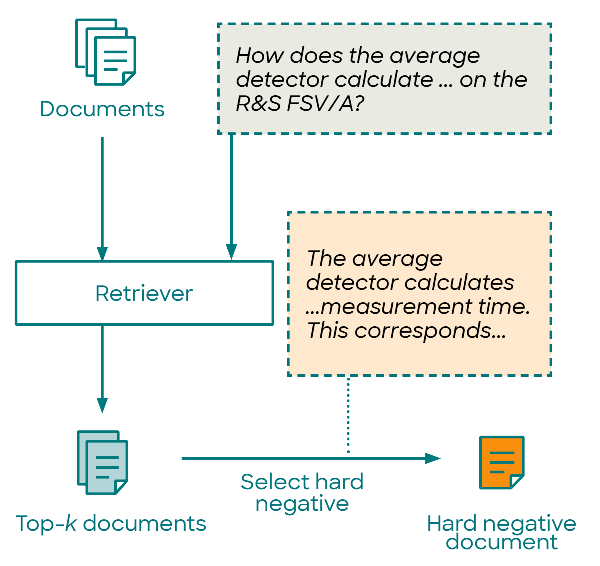
Incorporating expert/customer queries leads to more realistic generated questions: talk like your target group

Check example queries with experts to ensure realism

Enhancing Neural Understanding of Specialized Data: *Retriever Fine-Tuning*

Finding Hard Negatives

Pairs of **(query, relevant document)** are already enough for fine-tuning. Nevertheless, finding documents that are irrelevant, but hard to distinguish (so-called **hard negatives**), can further boost performance.



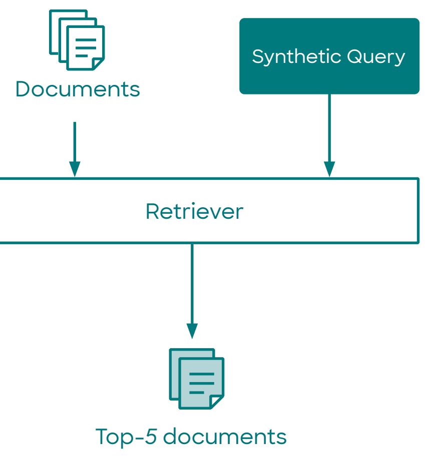
How can we select hard negatives?
- Ignore top-m results (m a number below k)
- Similarity score + margin below the relevant document score
- Similarity score below a threshold (check scores first!)

Which retriever should be used?
- We selected a retriever already fine-tuned on pairs (query, relevant document) because fine-tuning led to a more wide-spread score distribution (see "Evaluating the Retriever") that was easier to use for selecting hard negatives based on scores

Creating Test Sets

Retrievers are traditionally evaluated on a test set consisting of pairs (query, relevant documents), but this is typically not available for domain-specific problems. We aimed to **generate a test set with minimal manual effort**.

Creating Data for Labeling



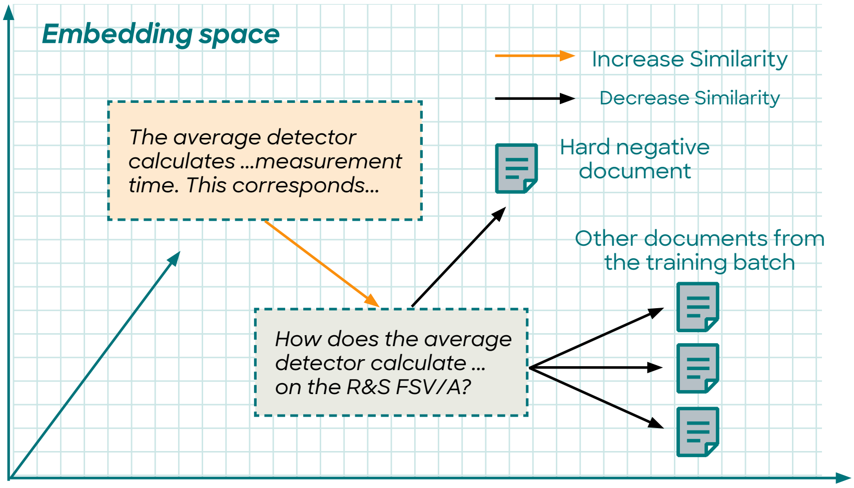
1. **Manual labeling:** we let domain experts label five documents per query for relevance to obtain a small manual test set (~20 queries in total)
2. **Label prompt creation and validation:** We designed an LLM prompt for automatic labeling of documents. Evaluation on the manually labeled test set showed an F1-score of ~0.8
3. **Automatic labeling:** we use the previously generated prompt to automatically label five documents per query for relevance (500 queries in total)

Besides evaluating our retriever on these two labeled test sets, we utilize **LLMs for evaluation**.
- **Explicit**: for a query and correct answer, evaluate the retriever for recall and precision
- **Implicit**: for a query, generate an answer based on the retrieved chunks and rate for correctness, leveraging the correct answer as a reference

Enhancing Neural Understanding of Specialized Data: *Retriever Fine-Tuning*

Training Methodology

We utilize **contrastive learning for retriever fine-tuning,** more specifically, the InfoNCELoss (sometimes also known as MultipleNegativesRanking loss)



- Training can be performed without hard negative documents, only using other documents in the batch as negatives **(in-batch negatives)**. This has already led to improvements
- **Staged training:** train with in-batch negatives first, and then add hard negative documents in the next fine-tuning round. Staged training performed slightly better than training in one step
- Choose your **batch size as large as possible** for better and more stable results

Info: Facts and Numbers

~450,000 documents in total

150,000 queries generated from ***70,000*** documents

70,000 hard negative documents mined

~$20 Cost to use GPT-4o-mini for query generation

1-2 hours Training time with batch-size 512 on four H100

Trained for one epoch

Frequently Asked Questions

Q: Is hard negative mining required?

A: In-batch negatives already improved results. Check whether this method already provides sufficient performance.

Q: How many queries do I need?

A: We already observed improvements with a query size much smaller than the 150.000 we used finally. Gradually increase the number and observe whether there are still improvements.

Q: Does it help to generate multiple queries per document?

A: In our case, this did not improve results further.

Q: How about generating negative queries for a given document?

A: This resulted in unstable learning and worse results in initial experiments.

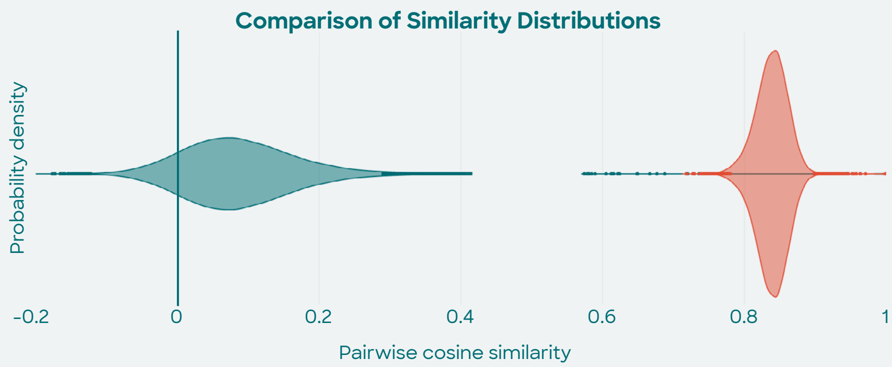
Enhancing Neural Understanding of Specialized Data: *Retriever Fine-Tuning*

Evaluating the Retriever

We compare the multilingual embedding model IntFloat/MultiLingual-E5-Large (**off-the-shelf**) before and after **staged fine-tuning**. All metrics consistently improve for both manual and automatically labeled test sets (for normalized discounted cumulative gain (NDGC)@10, Recall@10, Precision@10, Precision@1) and LLM-based evaluation. Note: "@10" for a metric means using the top 10 retrieval results for evaluation.

Understanding the Retriever

To better understand the similarity scores returned by the retrievers, we calculate the **pairwise cosine similarity** between documents within the corpus. This provides an overview of the score magnitudes and distribution we can expect. **The fine-tuned retriever** shows much more distributed scores, which can indicate more discriminative power. The vanilla model scores cluster around 0.86, making introducing score-based heuristics into the pipeline difficult.



Leading to Better Answers

Question:
How to obtain the frequency of all sweep points remotely?

Correct answer:
You have to query the stimulus data. The remote control command for this is CALC<Ch>:DATA:STIM.

Full Pipeline with Fine-Tuned Retriever

To obtain the frequency of all sweep points remotely, you can use the following SCPI command:

CALCx:DATA:STIM
Where x is the number of the channel.
This command queries the "Stimulus" data, which provides the frequency values for all sweep points...

Full Pipeline with Off-the-Shelf Retriever

Based on the provided documents, there are a few ways to obtain the frequency of all sweep points remotely:
For a linear frequency sweep, you can use the following approach:
Set the start and stop frequencies using [:SOURce<hw>]:FREQuency:STARt and [:SOURce<hw>]:FREQuency:STOP commands
...

Result:
The fine-tuned retriever successfully retrieves the correct information to answer the query.

From Manual to Automatic Evaluation: *Finding Reliable Automatic Metrics*

Efficient Evaluation Unlocks Fast Iteration Cycles

- **Reliable evaluation** is crucial for assessing the impact of RAG advancements and allows us to determine whether a specific new RAG pipeline component improves the overall system performance
- **Manual evaluation** is the most reliable way of assessment, but it is not always available and time-consuming, thus hindering assessments within fast iteration cycles
- **Automated evaluation utilizing LLMs' understanding and reasoning capabilities** is a highly dynamic field that opens up new possibilities
- We focus on metrics evaluating the entire pipeline here

Manual Evaluation by Experts

+ Gold standard: highest reliability
– Time-consuming
– Not always readily available

Automatic Evaluation

+ Fast
– Readily available
– Quality verification needed

Goal

Find automatic evaluation that correlates well with human judgment

Automated, LLM-Based Pipeline Performance Metrics

We initially evaluated our pipelines utilizing the LLM-based **Ragas answer correctness** [4]. After qualitatively comparing Ragas answer correctness scores on correct answers, we decided to propose our own **custom answer correctness based on LLM-as-a-Judge** as an alternative solution due to the following advantages and disadvantages.

Ragas Answer Correctness

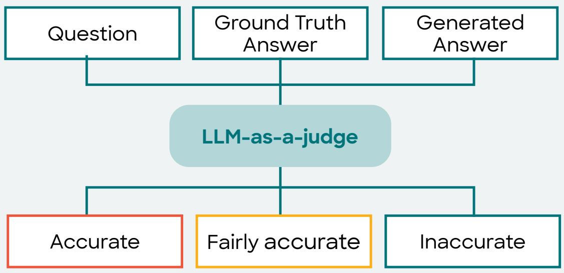
Explanation

- Measures the **accuracy of the generated answer** when compared to the ground truth answer
- Calculated as the **weighted average between answer semantic similarity and factual correctness.** Here, factual correctness quantifies the factual overlap between the generated answer and the ground truth answer
- Metric scale score ranging from 0 to 1

Disadvantages

- **Interpretability** of the score can be **challenging**
- **Tends to penalize LLM answers for verbosity,** even if judged as correct by an expert
- **Intransparent number of API requests and not customizable**

Custom Answer Correctness



Explanation

- Prompt an LLM to judge a generated answer as accurate, fairly accurate, or inaccurate [5]
- Inputs: question, ground truth answer, and generated answer
- Ordinal scale score taking on ranks -1, 0, or 1
- In this case, we used OpenAI GPT-4o-mini as LLM

Advantages

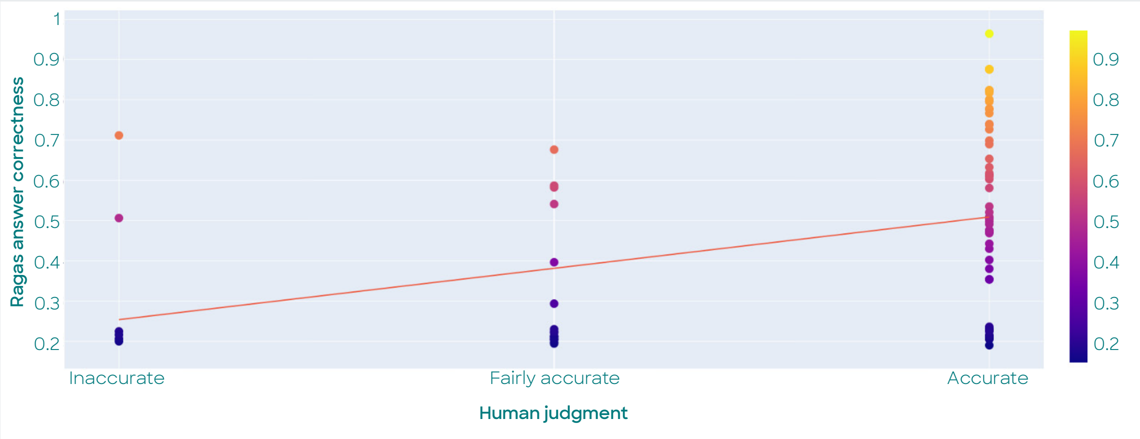
- Easy to compare with human judgment
- Full control over evaluation
- Can be advanced with prompt engineering and few-shot examples

From Manual to Automatic Evaluation: *Finding Reliable Automatic Metrics*

Correlation Between Automated and Human Evaluation Metrics

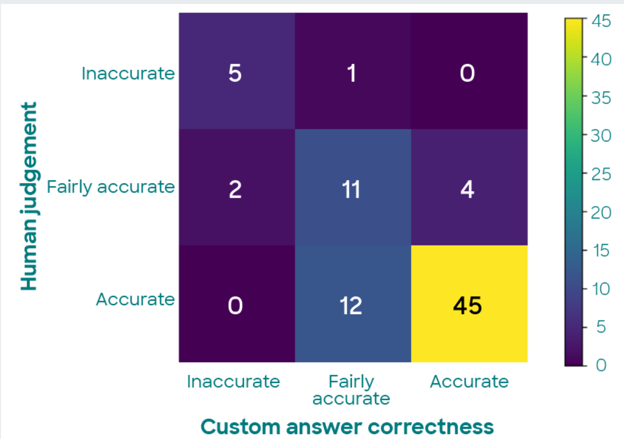
- We utilized the created QA set comprised of 79 questions and correct answers, together with generated answers and corresponding human judgment for our correlation study
- The **correlation between custom answer correctness and human judgment was higher** (Spearman's ρ = .65, p < .001) than between Ragas answer correctness and human judgment (Spearman's ρ = .40, p < .001)
- Due to identical categories to human judgment, custom answer correctness can be evaluated as a classification problem; F1 Score = .78 (method: weighted average)

Correlation: Ragas Answer Correctness vs Human Judgment



➡ For **Ragas answer correctness, inaccurate answers can lead to high answer correctness.** Moreover, accurate and fairly accurate answers exhibit a wide distribution of correctness scores, thereby complicating the interpretability of scores.

Confusion Matrix: Custom Answer Correctness vs Human Judgment



➡ **Confusions between inaccurate and accurate, which are the most severe misjudgments, did not occur for custom answer correctness**

Conclusion

- Custom answer correctness is more closely aligned with human judgment compared to Ragas answer correctness
- Enables evaluation in the absence of costly human expert feedback
- Easy interpretability and adaptability through prompt modification

Chapter 3: Learnings and Future Perspectives

Insights and Future Perspectives

What We Have Learned Along the Journey

Overall RAG Pipeline

- Out-of-the-box RAG pipelines generally perform well and provide a good starting point
- The effectiveness of a RAG pipeline depends on each component's strength - carefully assess the weak points to optimize the components that potentially lead to the highest improvements
- Fully optimized pipelines can effectively handle domain-specific questions, provided the necessary information is documented and accessible

Retrieval Improvement

- LLMs are in-domain query generators and can easily and cheaply generate realistic queries given in-domain prompts
- Fine-tuning on synthetic data is robust and reliably improves retrieval-specific metrics
- Fine-tuning leads to a much wider distribution of similarity scores that can be harnessed for heuristics
- Retrieval fine-tuning leads to improvement of the full pipeline, but the impact is less pronounced due to multiple influences such as hybrid retrieval, retrieving a large number of documents, reranking, and LLMs
- Retrieval worked best for English queries in this study

Visual Modality for Tables

- Vision pipeline as a fallback option is a viable solution to augment a text-based RAG pipeline
- The necessity for a vision pipeline depends on the strength of pipeline components: stronger components decrease the usage of the vision pipeline
- Vision pipeline alone can lead to hallucinations and inferior performance compared to text-based pipeline

Automated Evaluation Simulating Human Expert Feedback

- While out-of-the-box automatic evaluations such as Ragas are readily available, the provided answer correctness scores are often difficult to interpret and can require many LLM API calls that lead to intransparent costs
- LLM-as-a-judge paradigm that imitates the domain expert evaluation leads to easy interpretation of scores and higher correlations. Moreover, provided LLM reasons were sensible and further increased interpretability
- LLM-as-a-judge gives full control over evaluation and can be readily enhanced with few-shot learning

User Experience

- User education on system limitations is important to set realistic expectations and also helps to generate ideas and inputs for further pipeline enhancements
- Addressing specific user needs can significantly enhance user experience (e.g. dialogue interactions)
- Going past technical performance: discuss how the final product should look and operate

Conclusion

- Combining the strengths of in-domain and RAG experts, a tailored and high-performing RAG system can effectively be built in a short amount of time
- Depending on the strength of the text-based RAG pipeline, incorporating vision capabilities as a fallback option can deliver more accurate responses
- Automatic evaluation by imitating humans through LLM-as-a-judge provides high correlation, interpretability, and customizability for evaluation at scale
- The retriever, the single most important component for connecting company data with LLMs, can robustly be fine-tuned on synthetically generated data to understand complex domain language. In this context, realistic synthetic queries can be readily generated by LLMs in a cost and time-efficient way

Insights and Future Perspectives

Future Opportunities in RAG Technology

Boosting Retrieval

- Create and utilize metadata to improve retrieval and, consequently, answer generation
- Translate non-English queries to English for better retrieval results
- Leverage an agentic routing module to adjust the retrieval based on the type of questions, such as yes/no questions or open-ended questions
- For factual questions, tune the retrieval to prioritize specification files over manuals. Specification files contain much detailed information in a tabular format and demonstrate the preferred source
- Add a dictionary of domain-specific abbreviations and replace abbreviations in user queries. This might provide more context for retrieving the most relevant sources

Boosting Answer Generation

- Include multi-step reasoning agents to enhance specific types of questions, such as comparisons
- Explore large language models (LLMs) that offer extended context length and enhanced reasoning to identify relevant information from a larger context

Boosting User Experience

- Suggest follow-up questions at the end of each response for user guidance
- Provide the user with a measure of uncertainty in the generated answer to interpret trustfulness
- Carefully utilize users' search history to enrich the context of queries and improve the accuracy of answer identification

“The advanced natural language processing of Generative AI, integrated with RAG capabilities, is revolutionizing knowledge management by providing quick and effective access to internal resources for decision-making—similar to human interaction.”

– Bernhard Pflugfelder

Head of GenAI, appliedAI Initiative GmbH

References

[1] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey", March 2024

[2] P. Yu-Chun Chang, B. Pflugfelder, "Retrieval-augmented Generation Realized: Strategic & Technical Insights for Industrial Applications", June 2024

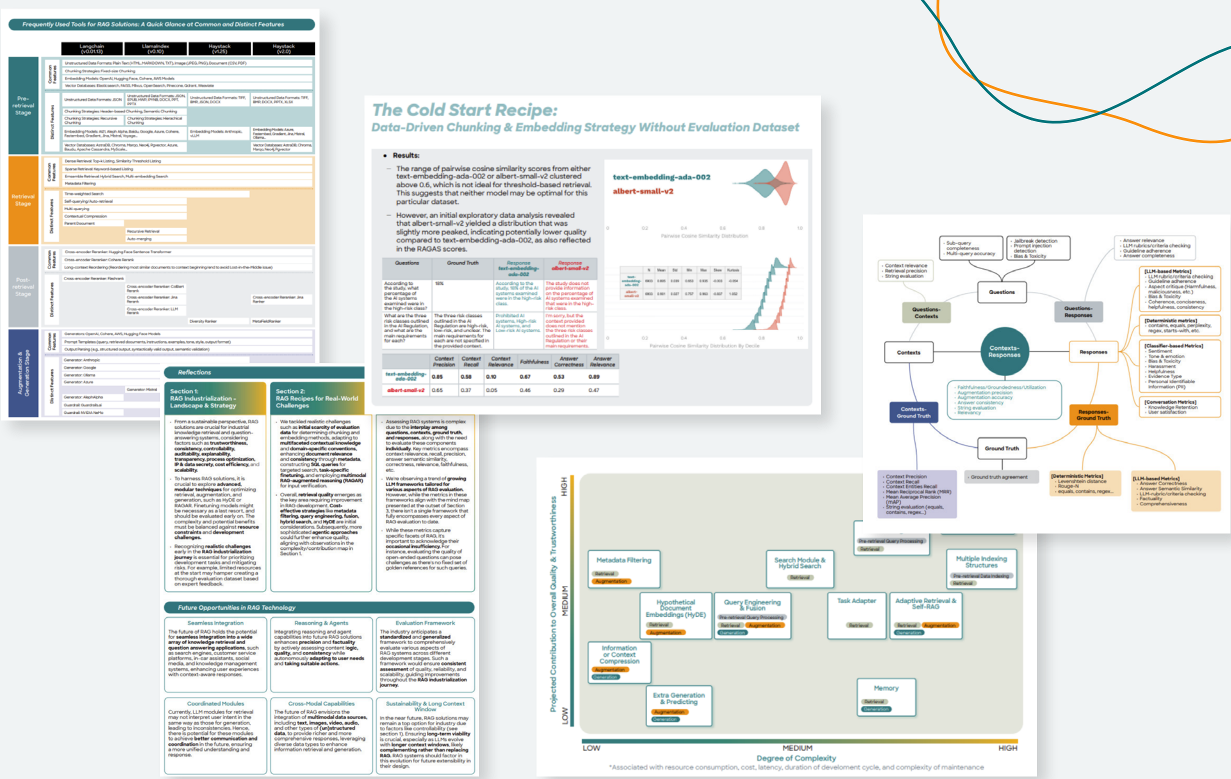
[3] S. Barnett et al., "Seven Failure Points When Engineering a Retrieval Augmented Generation System", IEEE/ACM 3rd International Conference on AI Engineering, April 2024

[4] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, "RAGAS: Automated Evaluation of Retrieval Augmented Generation", September 2023

[5] L. Zheng et al., "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena", 37th Conference on Neural Information Processing System (NeurIPS 2023) Track on Datasets and Benchmarks, June 2023

Do you want to dive deeper into RAG? Start your journey with our white paper *Retrieval-Augmented Generation Realized: Strategic & Technical Insights for Industrial Applications*

Our latest whitepaper on Retrieval-Augmented Generation (RAG) offers insights into the advancements and challenges of Retrieval-Augmented Generation (RAG) within the industry. It provides an analysis of industry demands, current methodologies, and the obstacles in developing and evaluating RAG. Additionally, our whitepaper aims to facilitate strategy development and knowledge exchange about practical use cases across various industrial sectors. The whitepaper is the result of extensive studies and discussions conducted with our internal teams and industry partners. It highlights RAG as a cost-effective technique that has significantly improved the trustworthiness and control of Large Language Model (LLM) applications over the past year.



Authors

Dr. Christian Geishauser
Generative AI Engineer
appliedAI Initiative GmbH
c.geishauser@appliedai.de

Christian Geishauser works as Generative AI Engineer at the appliedAI Initiative GmbH. His focus lies on large language models, agentic applications, and retrieval-augmented generation. Christian holds a PhD from Heinrich Heine University Düsseldorf where he focused on continual reinforcement learning for dialogue systems.

Johannes Birk
Generative AI Engineer
appliedAI Initiative GmbH
j.birk@appliedai.de

Johannes Birk is a Generative AI Engineer at the appliedAI Initiative GmbH. He works at the intersection of research and productization, focusing on NLP, LLMs, and RAG. In the past, Johannes leveraged NLP and text mining to automate processes in the insurance industry. He holds a master's in Cognitive Systems from the University of Ulm.

Bernhard Pflugfelder
Head of Generative AI,
appliedAI Initiative GmbH
b.pflugfelder@appliedai.de

Bernhard Pflugfelder works as Head of Generative AI at the appliedAI Initiative GmbH. Bernhard has 15 years of experience in the fields of Data Science, Natural Language Processing (NLP), as well as data and AI across different companies such as BMW Group or Volkswagen Group. He is renowned for his expertise especially in the field of AI in general, NLP and Generative AI in particular.

Contributors

Dr. Paul Yu-Chun Chang
Senior AI Expert: Foundation Models - Large Language Models,
appliedAI Initiative GmbH
p.chang@appliedai.de

Paul Yu-Chun Chang works as an Senior AI Expert specializing in Large Language Models at appliedAI Initiative GmbH. He has 10 years of interdisciplinary research experience in computational linguistics, cognitive neuroscience, and AI, and 6 years of industrial experience in developing AI algorithms in language modeling and image analytics. Paul holds a PhD from LMU Munich, where he integrated NLP and machine learning methods to study brain language cognition.

Dr. Ivan Rodriguez
Principal AI Engineer
appliedAI Initiative GmbH
i.rodriguez@appliedai.de

Ivan Rodriguez is a Principal AI Engineer at the appliedAI Initiative GmbH. He holds a Ph.D. in theoretical physics from the University of Florence and he has extensive experience in AI across both academia and industry. His work at appliedAI focuses on Reinforcement Learning, Bayesian ML, GenAI, and Computer Vision.

Philipp Joppich
Machine Learning Engineer
Rohde & Schwarz GmbH & Co. KG
philipp.joppich@rohde-schwarz.com

Philipp Joppich is a Machine Learning Engineer in the corporate Research and Development (R&D) division at Rohde & Schwarz. He specializes in Natural Language Processing (NLP) applications, particularly in deploying and integrating Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) systems. Philipp has also contributed to various machine learning projects across the company, focusing on time series data and tabular data processing. Philipp holds a Master's degree in Electrical Engineering from the Technical University of Munich.

Dr. Andrew Schaefer
Technology Coordinator, AI
Rohde & Schwarz GmbH & Co. KG
Andrew.Schaefer@rohde-schwarz.com

Andrew Schaefer is a passionate engineer working as technology coordinator for AI in the technology incubation team of the corporate Research and Development (R&D) division at Rohde & Schwarz. The technology incubation team is involved in multiple AI projects throughout Rohde & Schwarz, across divisions and in different levels of maturity. Additionally, Andrew chairs the cross-divisional solution group for data analytics. Andrew holds a PhD in communications engineering from the Technical University of Munich

Dr. Sebastian Husch Lee
Solution Engineering Tech Lead
deepset GmbH
sebastian.huschlee@deepset.ai

Sebastian Husch Lee has a strong background in machine learning, and a proven track record of driving innovation in both academia and industry. Currently, he is the Tech Lead of Solution Engineering at deepset, where he is responsible for developing and deploying NLP-powered solutions for enterprise clients. In this role, he applies his expertise of retrieval-augmented generative AI, and information retrieval to architect and implement scalable, production-ready solutions that drive business value for clients.

Laura Luckert
Senior Applied NLP Engineer
deepset GmbH
laura.luckert@deepset.ai

Laura Luckert is a Senior Applied NLP Engineer at deepset, where she implements NLP and Generative AI solutions with a strong focus on driving customer success and delivering measurable business impact. Specializing in retrieval-augmented generation and information retrieval, Laura combines her background in computational linguistics, computer science, and communication science with over 8 years of experience in data science, NLP, and generative AI across diverse industries to transform language technologies into practical solutions.

Johannes Steffens
Senior Director
Rohde & Schwarz GmbH & Co. KG
johannes.steffens@rohde-schwarz.com

Johannes Steffens is a Senior Director at Rohde & Schwarz. As a graduate engineer in electrical engineering with more than 30 years of professional experience, he heads various departments in the areas of spectrum and network analysis, joint research and development and is responsible for future-oriented pre-development. In this context, he also works on the use of strategically important competencies for R&S, such as the beneficial use of AI. He holds a degree in electrical engineering/communications engineering from the German University of Ilmenau.

About appliedAI Initiative GmbH

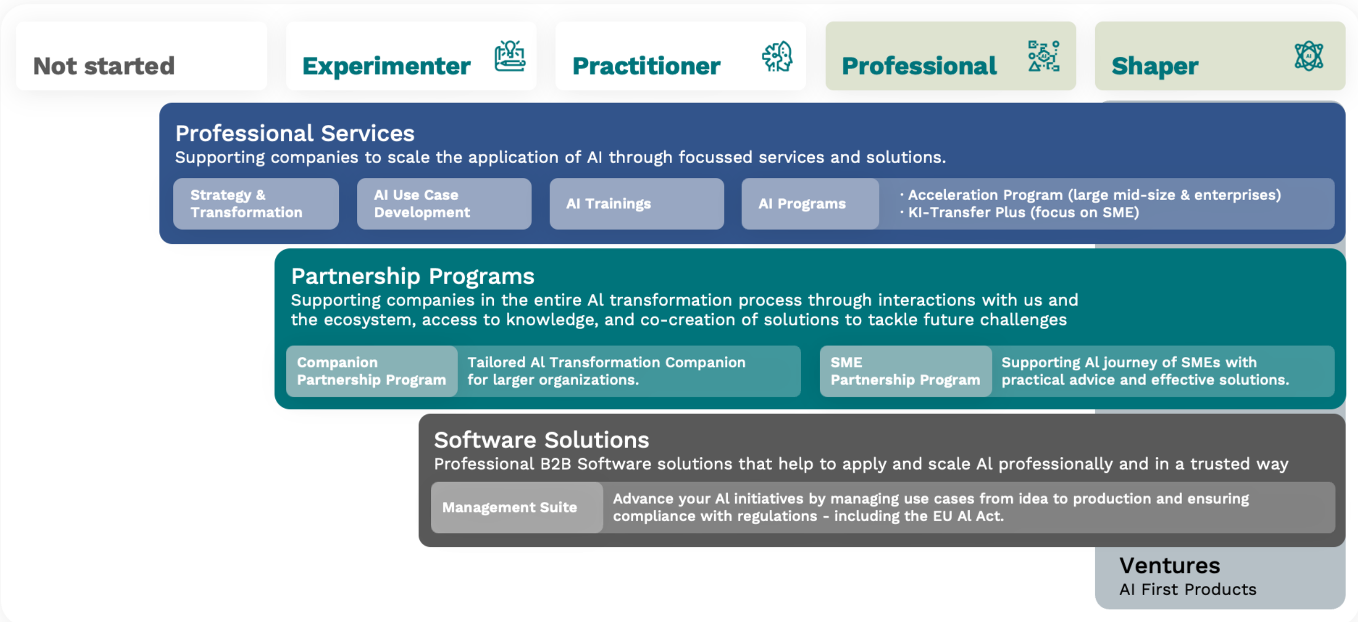
appliedAI is Europe's largest initiative for the application of trusted AI technology. The initiative was established in 2017 by Dr. Andreas Liebl as a division of UnternehmerTUM Munich and transferred to a joint venture with Innovation Park Artificial Intelligence (IPAI) Heilbronn in 2022.

At the Munich and Heilbronn offices, more than 150 employees pursue the goal of making European businesses a shaper in the AI era in order to maintain Europe's competitiveness and actively shape the future.

appliedAI holistically supports international corporations, including BMW and Siemens, as well as medium-sized companies in their AI transformation. This is accomplished through partnership-based exchange and joint knowledge building, comprehensive accelerator programs, and specific solutions and services, such as strategy consulting and Use-Case development.

For more information, please visit
https://www.appliedai.de/en/

We offer a unique set of offerings to help companies on their way to becoming AI shapers

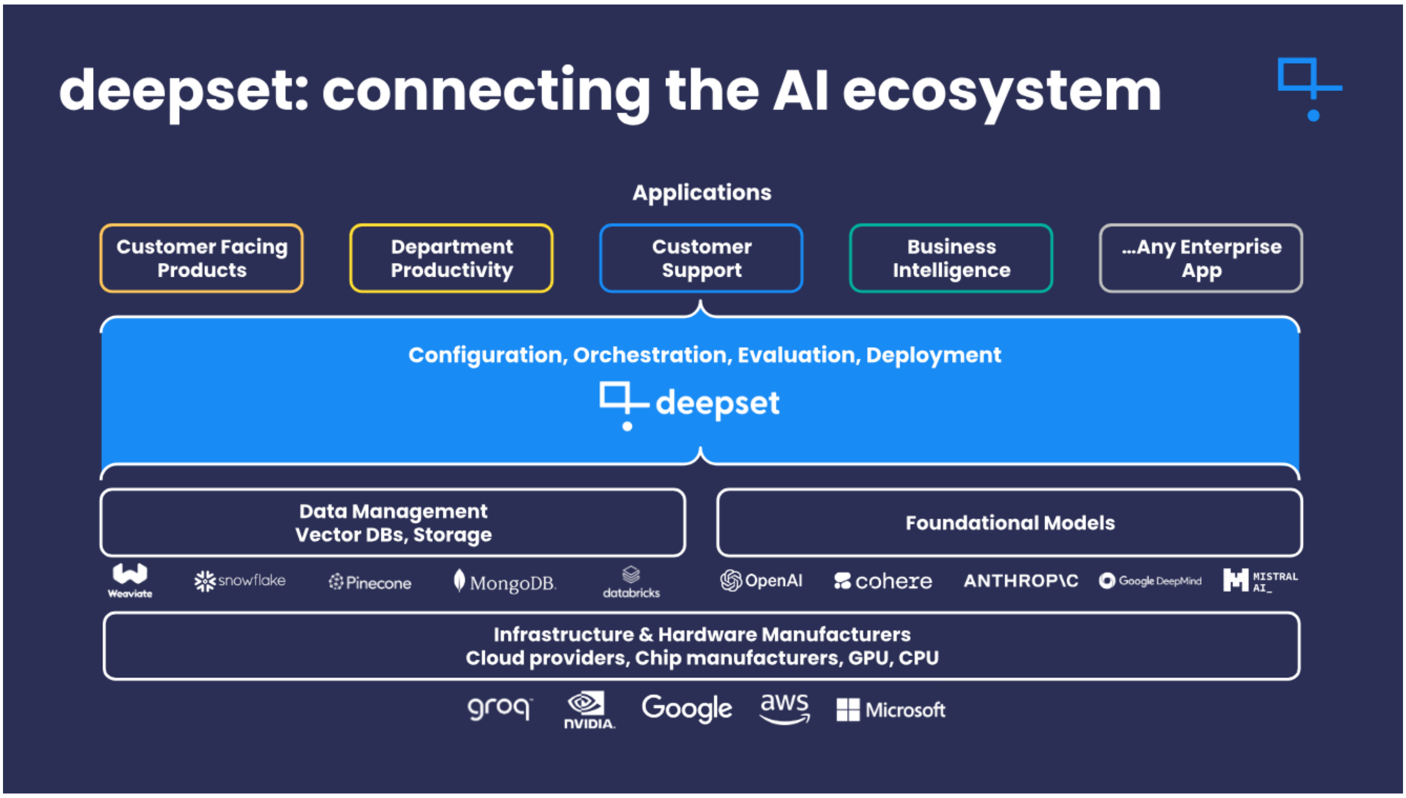


About deepset

deepset empowers organizations to build production-ready business applications with customized generative AI solutions. Founded in 2018 in Berlin, the company has grown into an international leader, driving AI adoption for enterprises across Europe and beyond.

deepset's commercial and open-source products, deepset Cloud and Haystack, are trusted by companies like Airbus, ZEIT, Siemens, and NVIDIA to accelerate the development of AI and LLM applications for critical use cases. With a global team, customer base, and community, deepset is committed to advancing AI technology, making it faster, more reliable, and fully tailored to meet the evolving needs of every industry.

For more information, please visit
https://www.deepset.ai/



About Rohde & Schwarz

Rohde & Schwarz is a global technology group striving for a safer and connected world. With its Test & Measurement, Technology Systems and Networks & Cybersecurity Divisions, the company creates tomorrow's innovations today. The company's leading-edge products and solutions empower industrial, regulatory and military customers to attain technological and digital sovereignty.

Innovation has been part of Rohde & Schwarz since the very beginning. The company founders Dr. Lothar Rohde and Dr. Hermann Schwarz were technological pioneers. With their hands-on entrepreneurial spirit, the two college friends entered the unexplored field of RF engineering. Ninety years later, the company is still pushing technological boundaries – as a successful shaper of cutting-edge technologies such as artificial intelligence (AI), 6G, cloud and quantum technologies. The privately owned company is known for stability and resilience. Independence is at the core of its entrepreneurial identity. The company finances its growth with its own resources.

Furthermore Rohde & Schwarz has an extensive sales and service network and is present in about 70 countries, primarily with its own subsidiaries. The company is headquartered in Munich, Germany. To maintain its high quality standards, flexibility and reliability, the company covers a significant share of the value chain in-house. Most of its products are developed in Germany.

Rohde & Schwarz has always seen the close proximity of R&D and production locations as an advantage. The employees at the German plants in Memmingen and Teisnach as well as the Czech town of Vimperk close to the German border manufacture most of the product range. Smaller plants in Singapore and Malaysia provide manufacturing capacity in close collaboration with R&D at the Asian headquarters. A number of subsidiaries in Germany and Europe ideally complement the group's solution portfolio in various market segments with their product range and expertise.

For more information about Rohde & Schwarz, please visit www.rohde-schwarz.com.

Acknowledgement

We express our sincere appreciation for the excellent work carried out by the authors, reviewers, and designer of this case study paper. Their great expertise, motivation, and dedication to every detail made this white paper an exceptional contribution to the AI community.

Furthermore, we extend our gratitude to all contributors for their valuable contributions throughout this joint case study. Their profound expertise and commitment have played a pivotal role in shaping the ideas, knowledge and results presented in this paper.

Our sincere thanks go to our appliedAI partner, Rohde & Schwarz, for their openness, commitment, and contributions to this case study. We highly appreciate their provision of the use case and data, as well as their willingness to share insights and results. Additionally, we are thankful for the time invested by various colleagues from the business units and central innovation units in evaluation cycles and discussions.

Lastly, we extend our gratitude to our partner, deepset, for their exceptional support and commitment throughout this case study. Their expertise, combined with their technology embodied in the deepset Cloud product, significantly enabled this collaboration.

The collective expertise, exchange, and dedication to advancing the knowledge in Generative AI and RAG were great inspirations throughout the process of creating this white paper.

How Do I Optimize the Dynamic Range of an FSW Signal and Spectrum Analyzer?

A RAG Use Case Study in Wireless Test & Measurement: Retrieval Fine-Tuning and Tables as Images

appliedAI Initiative GmbH

August-Everding-Straße 25
81671 München
Germany
www.appliedai.de