GRAG

A TRAINING-STUDY FOR FINE-TUNING SMALL LANGUAGE MODELS FOR GERMAN RETRIEVAL AUGMENTED GENERATION USE-CASES

hessian.AI

Foreword

Advancing German Language AI: The GRAG Model Suite

As organizations across industries increasingly adopt Generative AI, the need for specialized, language-specific solutions has become paramount. Addressing this need, **Avemio AG** and **hessian.AI** present GRAG (German Retrieval Augmented Generation), a comprehensive suite of AI models specifically optimized for the German language.

The GRAG suite represents a significant advancement in German-language AI capabilities, featuring:

- **Language Models**
- **Embedding Models**
- **Speech Recognition Model**
- **Training Data**

This suite of models represents a significant step forward in making advanced AI technology accessible and practical for German-language applications. By offering models that can run on consumer devices while maintaining high performance, GRAG enables organizations to implement AI solutions that are both powerful and practical.

The combination of language models, embedding models, and speech recognition capabilities provides a complete ecosystem for German language AI applications, from text generation to information retrieval and speech processing. This integrated approach ensures organizations can deploy comprehensive AI solutions while maintaining control over their data and processes.

Organizations implementing GRAG can expect:

- **Optimized performance for German language processing**
- **Resource-efficient deployment options suitable for various scales of operation**
- **Comprehensive toolset for building sophisticated AI applications & datasets**
- **Access to high-quality training data for further customization and development**

Executive Summary

The evolution of the GRAG (German Retrieval Augmented Generation) Suite emerged from a unique collaboration between Avemio AG and hessian.AI. This collaboration led to the development of extensive, high-quality datasets comprising over 3.1M training samples. These Samples were either synthetically enhanced or completely generated using larger open-source models, ensuring comprehensive coverage of German language nuances and use cases.

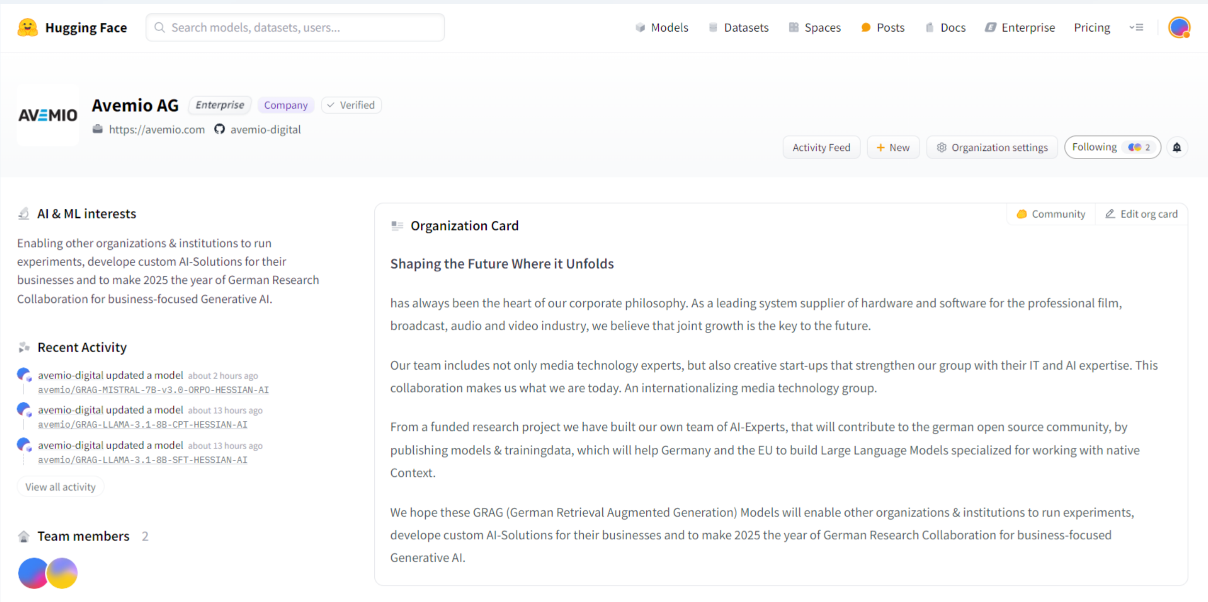
The result of this collaboration is the GRAG Suite, a comprehensive collection of AI models specifically optimized for German language applications. The suite includes:

- **4 Language Models (4B - 12B parameters)**
- **2 Embedding Models (0.35B & 0.5B parameters)**
- **1 Speech-to-Text Model (809M parameters)**

All LLM's were trained through a sophisticated three-stage process (CPT, SFT, and ORPO), leveraging the synthetically generated dataset. This approach has resulted in models that demonstrate superior performance in German language processing while maintaining efficient resource utilization suitable for consumer device deployment. The GRAG Suite represents a significant milestone in German language AI development, show-casing how strategic collaboration and innovative data generation approaches can advance the capabilities of open-source AI models for specific language applications.

We hope these GRAG Models will enable other organizations & institutions to run experiments, develop custom AI-Solutions for their businesses **and to make 2025 the year of German Research Collaboration for business-focused Generative AI.**

You can now access all Models & datasets here.



Intended Use of the LLM's

Improved Language Understanding:

We see overall improvements of the language quality & reliability to follow the trained instructions in a way that you can use these models as they are **when you stick to the prompt templates and only <u>slightly</u> adjust them to fit your needs.** These models are highly specialized and can be **adapted to your use-case by training them further** with Preference-Training with only a few thousand samples (we advise to use ORPO).

Pipeline Improvements:

Nearly all business-related RAG Use-Cases we have seen so far rely on domain specific knowledge & data. Because of that, we wanted to contribute to the development of these highly specialized systems by providing open-source models that are primed for German language understanding and generation. This will enable organizations to synthetically enhance their own data securely & locally to finetune specific models for their needs.

Bias, Risks and Limitations:

Like any base language model or fine-tuned model without safety filtering, it is relatively easy for a user to prompt these models to generate harmful and generally sensitive content. Such content can also be produced unintentionally, especially in the case of bias, so we recommend users consider the risks of applications of this technology. Despite lso the reliability was improved, many facts from GRAG-Models could not be true. So please ensure they get verified & checked. For that reason, we have included source citing directly into the model to have the possibility to check the used references to quickly evaluate the generated output.

Intended Pipeline:

We assume to use these models to first generate questions based on your domain specific context chunks. After that you should generate the context-based answer (Question + Context). With these triples (Question + Context = Answer) you can already do a basic finetuning of a custom LLM or Embedding Model. The next step would be to evaluate the performance of your model compared on a set of generated samples that you have <u>not</u> included in the training data. Depending on the intended use, you can directly use our prompt templates, adapt them slightly to your needs or train completely new ones.

Info: A Brief Introduction to RAG (Retrieval Augmented Generation)

The Challenges of using LLMs in Business-Context

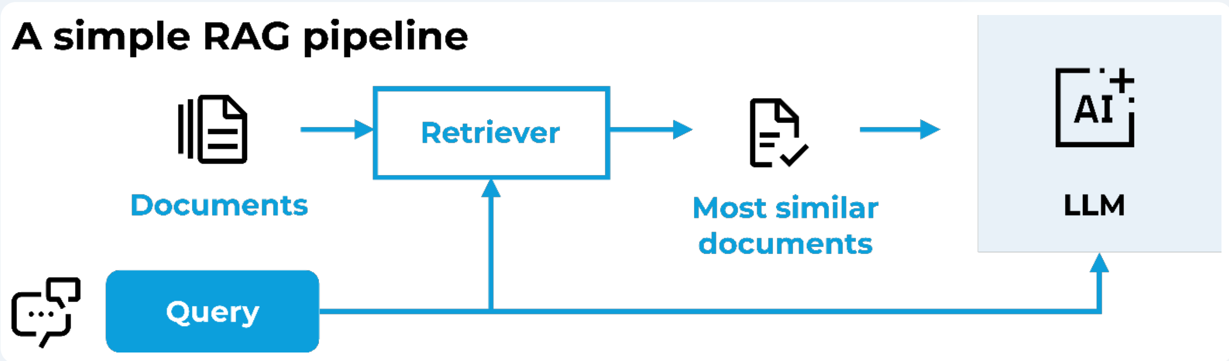
Natural language models provide exciting new opportunities through their ability to understand, create, and analyze text like humans do. However, these systems face several important limitations:

- Their “knowledge base” is restricted to what they learned during training
- They cannot access or understand private or company-specific information
- Their responses lack verifiable sources to support their claims

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 Company Information & LLMs

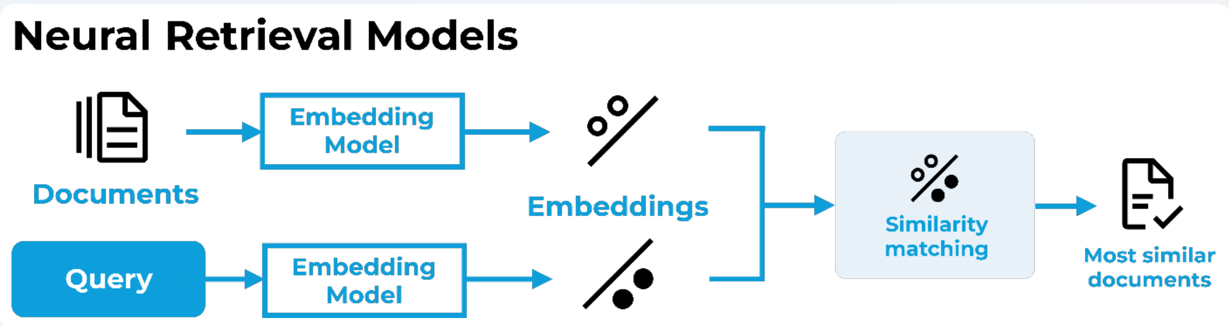
The RAG (Retrieval Augmented Generation) approach offers a solution by combining language models with retrieved contextual information from company documents, helping to overcome the shortcomings of using language models alone.



*

When given a question, RAG systems first search through their document collection to identify relevant information.

These retrieved documents are then provided to the language model along with the original question, enabling it to generate responses that incorporate specific organizational knowledge.



*

Unlike traditional search systems that match words exactly, neural retrieval systems understand the meaning behind both questions and documents. These neural systems find matching documents through two main phases:

- First, the system converts both the question and all documents into numerical sequences (known as embeddings)

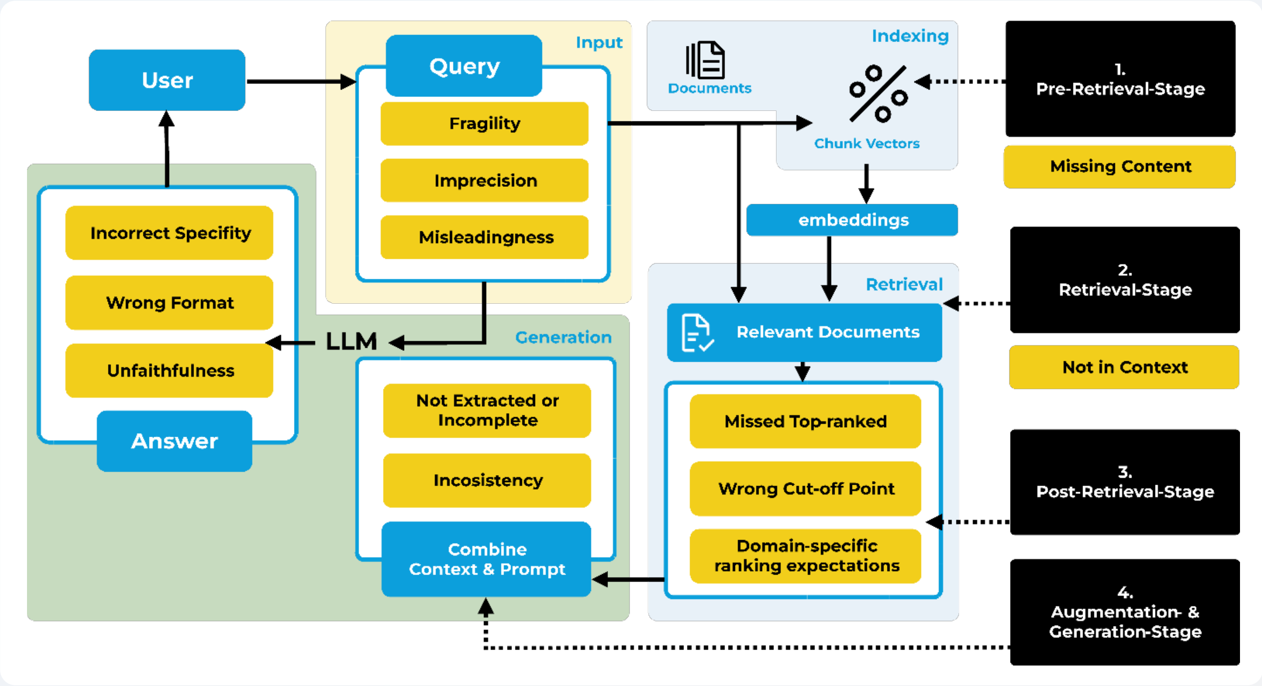
*Visualization adapted from https://www.appliedai.de/uploads/files/trustworthy-rag-in-wireless-test-measurement-retrieval-fine-tuning-and-tables-as-images/How-Do-I-Optimize-the-Dynamic-Range-of-an-FSW-Signal-and-Spectrum-Analyzer.pdf

- Then, it compares these numerical representations to find documents that are most conceptually similar to the question

A specialized neural network, called an embedding model, handles the creation of these numerical representations.

Info: General Challenges in a RAG System

- Questions from users can be vague, potentially misleading, and often rely on unstated background information or knowledge.
- The preparation phase involves processing and organizing documents but faces challenges like incomplete content extraction and dealing with text converted from OCR, which often needs cleanup. Additionally, getting complete context sometimes requires converting speech and video content into text format.
- Finding the correct supporting information is a crucial step that must be carefully managed to ensure accuracy.
- After retrieving documents, the process of refining and ranking results can be problematic due to inappropriate cutoff thresholds and failing to account for industry-specific ranking requirements.
- The final steps of enriching the context and creating responses can be compromised by contradictions in the retrieved documents and the generation of responses that are either inaccurate or fail to faithfully represent the source material, even when the retrieved information is correct.



*

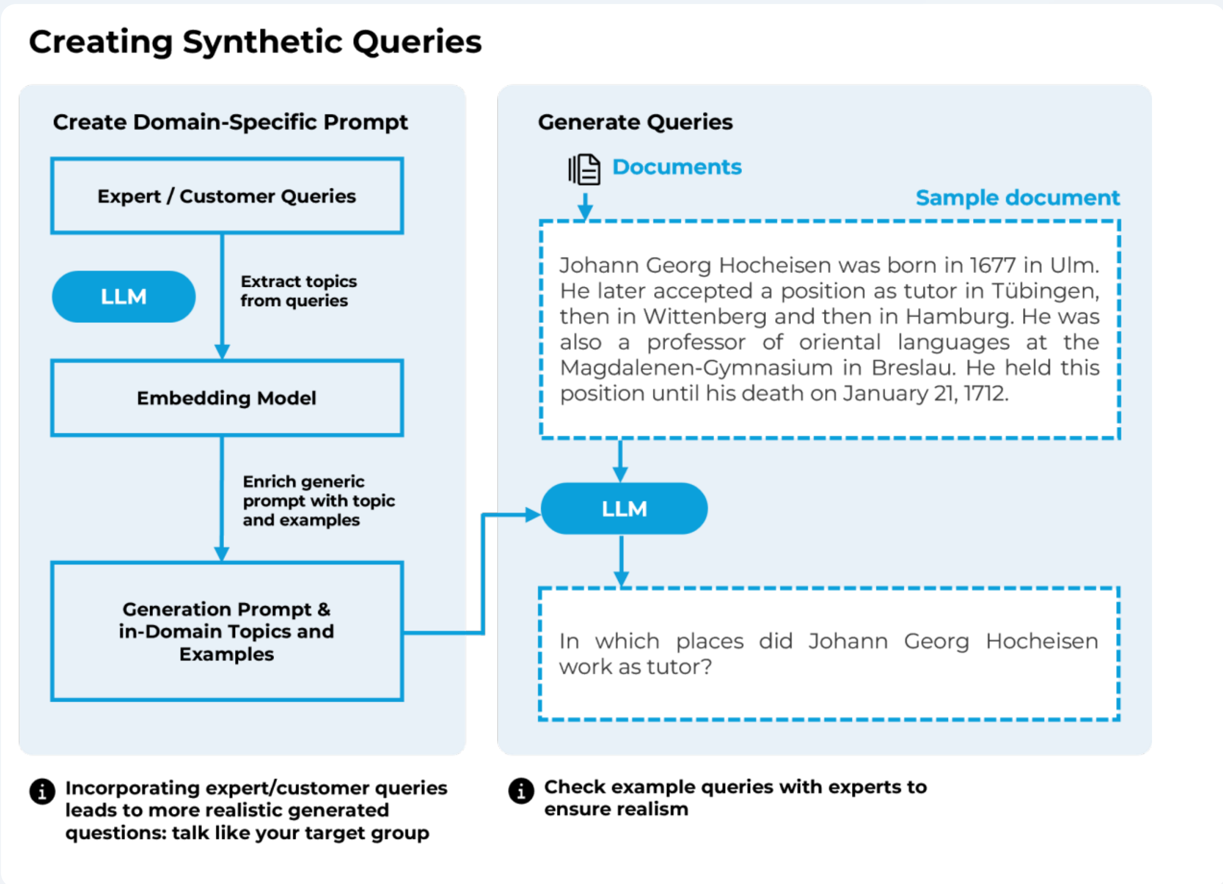
*Visualization adapted from https://www.appliedai.de/uploads/files/trustworthy-rag-in-wireless-test-measurement-retrieval-fine-tuning-and-tables-as-images/How-Do-I-Optimize-the-Dynamic-Range-of-an-FSW-Signal-and-Spectrum-Analyzer.pdf

Enhancing Neural Understanding of Specialized Data:

Finetuning for RAG-Use-Cases

Finetuned LLMs and Embedding Models are crucial for successful RAG application when working with Smaller Open-Source Models, but Base-Models are often trained on general and english knowledge, limiting their effectiveness for company or domain-specific adoption. Customizing these models through fine-tuning can significantly boost the performance in terms of reliability and answer correctness for domain-specific RAG applications, but acquiring domain-specific data for fine-tuning was before 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 labelling. Our goal is to enable any organization to generate synthetic training data for their specific use-case & domain easily by utilizing our primed models.

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



*

*Visualization adapted from https://www.appliedai.de/uploads/files/trustworthy-rag-in-wireless-test-measurement-retrieval-fine-tuning-and-tables-as-images/How-Do-I-Optimize-the-Dynamic-Range-of-an-FSW-Signal-and-Spectrum-Analyzer.pdf

Training Approach & Methods

Optimizing Base Models for Context-Dependent RAG Applications

Our training approach follows a carefully structured three-phase methodology designed to optimize base language models specifically for Retrieval Augmented Generation (RAG) applications in German language. The primary goal is to develop models that excel at generating responses that rely solely on provided context rather than their pre-trained knowledge. This approach is particularly crucial for business applications where accuracy and verifiability of information are paramount.

Phase 1 | Continued Pre-Training (CPT): The process begins with Continued Pre-Training, where base models are exposed to extensive German and English language examples (approximately 420,000 German and 200,000 English samples). These examples focus on three core competencies: context-based question answering, structured reasoning, and summarization. This phase establishes the foundational patterns for processing and responding to queries while maintaining strong ties to provided context. The deliberate emphasis on German language content expands the model's German vocabulary and capabilities while maintaining sufficient English language competency.

Phase 2 | Supervised Fine-Tuning (SFT): Building on the CPT foundation, the Supervised Fine-Tuning phase introduces more specialized tasks with structured output requirements. Using primarily synthetically enhanced data from German Wikipedia, organized in a knowledge graph structure, the models learn to handle specific task types ranging from classification and information extraction to time-aware question answering and function calling. This phase reinforces the model's ability to work within defined constraints while maintaining strict adherence to provided context.

Phase 3 | Odds Ratio Preference Optimization (ORPO): The final phase focuses on optimizing the model's reasoning capabilities and output quality. Through structured tasks emphasizing systematic problem-solving, constrained content generation, and quality control, ORPO ensures that the model's responses are not only accurate but also well-reasoned and appropriately formatted. This phase is crucial for developing models that can handle complex queries while maintaining strict context dependence.

The cumulative effect of this three-phase approach is a model specifically optimized for RAG applications, capable of:

- **Generating responses that rely strictly on provided context**
- **Processing and synthesizing information from multiple sources while maintaining accuracy**
- **Providing clear references and citations for all information used**
- **Handling temporal aspects of information appropriately**
- **Delivering responses in consistent, structured formats suitable for business applications**

This methodology ensures that the resulting models are particularly well-suited for enterprise applications where information accuracy, verifiability, and contextual relevance are critical. By prioritizing context-dependent reasoning over general knowledge application, the models provide more reliable and traceable outputs for business-critical applications.

Training Overview for Large Language Models (LLMs)

Approximate Training Time Based on Model Size and Hardware:

Model Name	Model Size	Hardware	Approximate Training Time
PHI-3.5-MINI-Instruct	3.5 billion parameters	8 A100 GPUs	9 days
MISTRAL-7B-v3.0	7 billion parameters	16 A100 GPUs	14 days
LLAMA-3.1-8B	8 billion parameters	16 A100 GPUs	14 days
NEMO-12B	12 billion parameters	32 A100 GPUs	18 days

Distributed Training Techniques

Distributed training is critical for reducing the time required to train LLMs. Below are the methods included in our training pipeline:

Data Parallelism

- The dataset is partitioned across multiple devices.
- Each device holds a replica of the model.
- Gradients are aggregated from all devices to update the model.

Model Parallelism

- The model is split among multiple devices.
- Each device computes a specific subset of operations.

Pipeline Parallelism

- Individual layers of the model are assigned to different devices.
- Activations are passed sequentially between devices, resembling a pipeline.

Hyperparameter Optimization

Advanced Optimization Algorithms

Optimized learning rate schedules were used to improve model convergence and generalization. Below are the techniques applied:

Learning Rate Schedule	Description
Step Decay	Reduces the learning rate by a fixed factor at specified intervals.
Cosine Annealing	Smoothly reduces the learning rate following a cosine curve.
Linear Warmup	Gradually increases the learning rate from a small value to the initial rate.
Cyclical Learning Rates	Alternates the learning rate between upper and lower boundaries in cycles.

Combined Approach

By integrating distributed training strategies with advanced learning rate schedules, we achieved:

- **Significant reduction in training time.**
- **Effective model convergence.**

CPT – Continued Pre-Training

Our CPT (Continued Pre-Training) approach is designed to enhance language models' ability to perform specific tasks through structured instruction-based learning. Drawing inspiration from ["Instruction Pre-Training: Language Models are Supervised Multitask Learners"] our methodology focuses on priming base models with semi-structured examples to improve their performance across three key tasks. **Our training dataset comprises approximately 420,000 German language samples and 200,000 English examples**, with the deliberate emphasis on German content aimed at expanding the model's German language vocabulary and capabilities.

Context-Based Question Answering

This task trains models to generate accurate responses by considering both the question and its accompanying context. For example, when analyzing cancer counseling center benefits, the model learns to extract and synthesize relevant information from provided context to formulate comprehensive answers. The training examples follow a clear structure:

Question > Context > Context-based Answer

Structured Reasoning

The reasoning task develops the model's ability to break down complex problems and arrive at solutions through systematic thinking. Training examples present problems with clear subheadings (Task, Approach, Solution) to encourage structured analysis. As shown in the music festival scheduling example, this format helps the model learn to consider multiple constraints and develop logical solutions step by step.

Problem > Approach > Solution

Intelligent Summarization

The summarization task teaches models to distill complex information into clear, organized summaries while preserving key details. Training examples demonstrate how to transform detailed explanations into well-structured bullet points or concise summaries.

Full-Text > Bullet-Points Summary

The CPT methodology prioritizes:

- **Clear task structures with explicit sections and subheadings**
- **Diverse example types across different domains**
- **Progressive complexity in reasoning tasks**
- **Balance between technical accuracy and clarity**
- **Consistent formatting patterns that help models recognize task types**
- **Strong emphasis on German language proficiency through greater German sample representation**

This approach aims to enhance model performance through carefully curated examples that demonstrate not just what to learn, but how to approach different types of problems systematically. The significant proportion of German language samples (roughly 68% of the dataset) ensures robust German language capabilities while maintaining sufficient English language competency for multilingual applications.

CPT-Task: Question Answering

Text:

What advantages does the financing of cancer counseling centers by statutory health insurance companies bring for cancer patients and the health care system as a whole?

Context:

'context': [

{'id': 160440, 'source': 'In November 2015, Spahn published a book entitled \'Ins Offene\', in which Boris Palmer, Julia Klöckner, Klaus von Dohnanyi, Wolfgang Ischinger, Markus Söder and Markus Kerber, among others, present their views on the refugee crisis in Europe since 2015. In the book, Spahn calls for a more open discourse on migration policy, speaks of a "disruption of the state" and in this context criticized Chancellor Angela Merkel's refugee policy.'},

{'id': 160439, 'source': 'Spahn had commissioned the ver.di union and the German Hospital Association to develop an instrument for assessing nursing staff. The ver.di union presented this with the PPR 2.0 in January 2020, which disappeared into a drawer in the ministry. At a press conference (October 11, 2021) he promised to present an effective staff assessment instrument by the end of 2024. The Walk of Care presented these and other demands, training regulations, voting rights in the G-BA, to the minister for a whole year from 2020 to 2021, without a response. Hospital strikes by nursing staff over relief collective agreements took place in 2021 and previously; such a concluded collective agreement then has the same authority as a law, which is no longer subject to political interference (agreement between the collective bargaining parties). Most recently in 2021 at the Nursing Council, Spahn warned of the poor level of organization in nursing and especially in geriatric care.'},

{'id': 160441, 'source': 'In July 2017, he told the world that those who come "to us" from the Arab cultural area are "often characterized by repressed sexual morality, by the lack of equal rights for women, the rejection of Jews or gays." Immigrants must relearn how to live in an open and liberal society, otherwise German society runs the risk of becoming "more anti-Semitic, homophobic, macho and violent" than it has been so far. This statement was praised by Alexander Gauland (AfD) in an open letter. A report in the Münchner Merkur pointed out that there were also Germans who opposed the equality of sexual minorities, such as supporters of the "Demo for All", which is supported by the AfD, and accused Spahn of profiling himself on the "right side of the Chancellor".'},

{'id': 160438, 'source': 'In November 2020, after talks with the Ministry of Labor and Social Affairs, Spahn unilaterally decided that statutory health insurance companies would have to finance the approximately 250 cancer counseling centers in Germany.According to a report by the Spahn Ministry, the average annual cost of such a counseling center is 161,500 euros. This would require around 40 million euros annually. Counseling is free for cancer patients.'}

]

The financing of cancer counseling centers by statutory health insurance companies means that cancer patients receive free counseling and the health system as a whole benefits from improved care and support for patients.

CPT-Task: Reasoning (Subsets DE + EN)

Text:

Task:

A local music festival plans to host three emerging Afrikaans artists. Each artist has different requirements for performance time, stage and technical equipment. Artist A requires the main stage and wants to perform at 18:00. Artist B requires the side stage and wants to perform at 20:00 but can only perform with special lighting equipment. Artist C can perform on any stage but does not want to perform before 19:00 and needs additional microphones. However, the festival only has a certain amount of technical equipment and a fixed order for performances. The technical equipment must not be overused, and no two artists can perform at the same time.

Approach:

To accommodate the artists' requirements, performance times and stage availabilities must be coordinated so that all artists meet the necessary requirements. Artist A can only perform on the main stage at 18:00. Artist C can only perform at 19:00, which means he cannot perform at the same time as artist A. Artist B requires the side stage and relies on special lighting equipment that may need to be tested beforehand. The order of performances must be set to accommodate technical requirements and stage availability without causing conflicts.

Solution:

Artist A should perform on the main stage at 6pm, followed by Artist C on the side stage at 7pm, as he can perform on any stage and does not want to perform until after 7pm. Artist B should then perform on the side stage at 8pm, after the special lighting equipment has been prepared for his performance. This order meets all of the artists' requirements and avoids conflicts in performance times and technical needs.

CPT-Task: Summarization

Text:

A spherical capacitor is an electrical capacitor that consists of two concentric, mutually insulated metallic spherical surfaces. ε is the electric field constant. ε is the dielectric constant, which is equal to 1 in a vacuum. For an infinitesimally small spherical shell between "R" and "R", the well-known relationship of the plate capacitor applies to the infinitesimally small reciprocal of the capacitance: If formula_11 is, one can approximately set formula_12 and obtain formula_13. The capacitance is then practically only determined by the radius of the inner sphere. This approximation also describes the capacitance of a free-standing sphere (also called a spherical electrode), since the counter electrode is very far away here (formula_14 and thus formula_15). The radius of such a spherical electrode in a vacuum used to be used as a unit of measurement for capacitance with the following equivalence: For a spherical capacitor, it is assumed that the two electrodes are oppositely charged with the charge formula_17 and formula_18. These charges are located as surface charges on the inward-facing spherical surfaces. The charge density can then be written as formula_19 The vector of the electric field between the two capacitor shells consists only of the radial component formula_21 due to the spherical symmetry. This can be calculated using the following formula: The field is not homogeneous but depends on the distance formula_24 to the center of the capacitor. There is no electric field inside the electrodes and outside the capacitor. The electric potential is a scalar field that only depends on formula_24 and is calculated, except for an additive constant, as formula_26. This integral can be determined section by section:

Summary as bullet points:

- A spherical capacitor consists of two concentric, insulated spherical surfaces, and the electric field constant is represented by the dielectric constant ε, which has the value 1 in a vacuum.

- For an infinitesimally small spherical shell between two radii, the capacitance relationship is used in a similar way to a plate capacitor, whereby the capacitance depends mainly on the radius of the inner sphere.

- This approximation is also applied to free-standing spheres, whose capacitance in relation to the size of the spherical electrode in a vacuum was previously used as a unit of measurement.

- In the spherical capacitor, the two electrodes are oppositely charged, which allows the surface charges on the spherical surfaces to be represented.

- Due to the spherical symmetry, the electric field vector between the capacitor shells only has a radial component and is dependent on the distance to the center of the capacitor.

- There is no electric field inside the electrodes and outside the capacitor, and the electric potential only depends on the distance, where it can be calculated by an integral.

SFT – Supervised Fine-Tuning

The SFT tasks represent a focused approach to enhance model capabilities through specialized RAG examples. Most of these tasks were developed using synthetically enhanced data derived from the **German Wikipedia, accessed through Cohere's prepared dataset on HuggingFace** (licensed CC-BY-SA 4.0). This data was structured in a training knowledge graph where Question-Answer nodes were connected to both relevant and irrelevant Context nodes from the **same Wikipedia page**, creating a rich and challenging network of relationships for training. The only exceptions are the function calling dataset, which was derived and extended from **Salesforce's XLAM Function calling dataset** by including function call results and final answer generation, and the reasoning task which synthetic generation was inspired by the Paper from Tencent **"Scaling Synthetic Data Creation with 1,000,000,000 Personas"**, to generate a diverse set of reasoning tasks across various domains.

Classification in JSON-Format

Training the model to categorize inputs according to predefined schemas, with structured JSON output for consistent data handling. This enables automated document routing, customer inquiry categorization, and standardized data extraction from unstructured business documents.

Schema > Context > JSON

Information Extraction and Recall

Teaching the model to identify answerable questions from given contexts and provide precise references to source information. This capability supports legal document analysis, compliance verification, and efficient technical documentation searches.

Question > Context > Relevant-Chunk Reference

OCR Text Correction

Developing the ability to clean and correct OCR-derived texts, fixing character recognition errors and ensuring proper grammar and spelling. This is crucial for processing scanned business records, digitizing paper-based contracts, and converting handwritten forms into accurate digital text.

OCR-Text > Corrected Text

Multiple-Reference Question Answering

Training the model to answer questions using multiple context sources while clearly citing specific references. This supports complex customer support scenarios, policy interpretation, and comprehensive product research combining multiple sources.

Questions > Context > Answer 1 + Reference & Answer 2 + Reference

Time-Aware Question Answering

Teaching the model to consider temporal aspects when providing answers, including appropriate disclaimers for dated information. This is essential for market analysis, policy compliance, and tracking documentation versions.

Question > Context > Context-based Answer + Time-Difference + Reference

Question Generation

Developing the ability to formulate clear, relevant questions about given text passages while maintaining proper naming conventions. This supports automated quality assurance, training material development, and customer FAQ generation.

Context > Question that can be answered by the Context

Structured Reasoning

Training the model to break down complex problems and provide step-by-step analysis before reaching conclusions. This enables sophisticated decision-making in project management, risk assessment, and strategic planning.

Problem > Repeat important Info > Approach > Solution

Context Relevance Assessment

Teaching the model to evaluate and select the most relevant context for answering specific questions. This improves enterprise knowledge base searches, customer query routing, and targeted content selection.

Question > Context > Most relevant Context-Chunk

Summarization Tasks

Developing the ability to distill essential information from longer texts while maintaining accuracy and coherence. This facilitates executive brief generation, meeting minutes automation, and contract summary creation.

Summarization Instruction > Context > Summarization (full or bullet-points)

Function Selection and Calling

Training the model to identify and properly utilize appropriate tools and functions based on user requests. This enables automated workflow orchestration, system integration, and business process automation.

Question + Tools > Tool-Call 1 > Tool-Response > Tool-Context-based Answer

This comprehensive set of SFT tasks ensures the model develops robust capabilities across a wide range of practical applications while maintaining consistent output formats and clear communication patterns. Each task type has been carefully designed to address specific business needs while maintaining high standards of accuracy and reliability, making them valuable tools for organizations looking to enhance their information processing and knowledge management capabilities.

Task Instruction Format

The implementation of these SFT tasks follows a carefully structured format designed for consistency and clarity. Each task begins with comprehensive system instructions often wrapped in XML tags that meta-define expected inputs, outputs, constraints, and example interactions. This standardization enables clear communication between the model and users while ensuring reliable results.

The context information utilized in these tasks is provided in a standardized JSON structure, including unique identifiers, source text, timestamps where relevant, and task-specific metadata. This format was specifically chosen to allow seamless integration with retrieved data from RAG systems, eliminating the need for additional formatting steps in production environments.

Source references are handled through a consistent system of numerical indices for context references, JSON-formatted citation markers, and clear time-difference notifications when temporal aspects are relevant. This systematic approach to referencing ensures traceability and reliability in the model's responses.

The implementation of these tasks within RAG systems can significantly improve organizational efficiency by reducing manual processing time, ensuring consistency in information handling, improving accuracy in data extraction and analysis, and enabling faster decision-making through better information access.

SFT-Task: Classification JSON

System:

Extract the category and type of the following statement in JSON format for creating a summary from these lists of available values:

Categories: request, question, instruction, request, work instruction

Types: full, bullet-points

Respond in this format: {'category': 'selected category', 'type': 'selected type'}

If none of the categories match, respond with an empty string as the value for the category {'category': '', 'type': 'selected type'}

User:

Summarize the provided passage of text in complete sentences. Focus on capturing the key points about publication titles, recording audio tracks, and the meaning of labels. Be sure to use clear and concise language to present the information clearly.

Assistant:

{'category': 'work instruction', 'type': 'full'}

SFT-Task: Extraction-Recall

System:

Which of the following questions can be answered using the context section?

If the information to answer the questions is not contained or derivable in the context, you say 'These questions cannot be answered using the available context.'

Here is an example of a query and how you should respond to it:

<Query> Context: This is where the context section would go. </Query>

Questions:

'questions':

```
[
        {'id': 500017, 'question': 'Here would be a question that can be answered with the context?'},
        {'id': 500020, 'question': 'Here would be an irrelevant question that CANNOT be answered with the context?'},
        {'id': 500018, 'question': 'Here would be an irrelevant question that CANNOT be answered with the context?'},
        {'id': 500019, 'question': 'Here would be an irrelevant question that CANNOT be answered with the context?'}
]
```

<Your answer> The question with the ID 500017 can be answered using the context section. Here would be a question that can be answered with the context? </Your answer>

User:

Context:

Beria now became First Deputy Prime Minister and Minister of the Interior in Georgi Malenkov's cabinet. Beria brought about the renewed merger of the Ministry of the Interior and the secret service in order to gain as much power as possible at the head of the organization thus created in the struggle for succession against Malenkov and Nikita Khrushchev. Just a few days after Stalin's death, Beria ordered the first steps towards de-Stalinization, such as the release of the accused arrested in the course of preparing the doctors' trial and the arrest of those responsible in the secret service, as well as the ban on torture.

Questions:

```
'questions': [
        {'id': 250001, 'question': 'What steps did Beria take shortly after Stalin's death to strengthen his influence in the struggle for succession and to reduce tensions in the country?'},
        {'id': 250003, 'question': 'What was the charge against Beria in the secret trial in December 1953?'},
        {'id': 250004, 'question': 'When was Lavrenti Beria shot?'},
        {'id': 250005, 'question': 'Who was Lieutenant General Vadim Volkovitsky and what role did he play in Beria's execution?'},
        {'id': 250007, 'question': 'What reasons did the Military Collegium of the Supreme Court of the Russian Federation give for refusing to rehabilitate Beria?'},
        {'id': 250006, 'question': 'What was the reason that in 1954 the editorial team of the Great Soviet Encyclopedia issued an urgent recommendation to its subscribers to remove the portrait and the accompanying text about Beria from the fifth volume of the encyclopedia with scissors and a razor?'},
        {'id': 250008, 'question': 'What evidence is there for Beria's claim that he poisoned Stalin?'},
        {'id': 250009, 'question': 'What role did Nina Gegechkori play in Beria's life?'},
        {'id': 250002, 'question': 'What motivated Beria to campaign in the Politburo for German reunification on the basis of neutrality and democracy?'}
]
```

Assistant:

The question with ID 250001 can be answered using the context section. What steps did Beria take shortly after Stalin's death to strengthen his influence in the struggle for succession and reduce tensions in the country?

SFT-Task: OCR-Correction

System:

Please correct these OCR-derived texts into a grammatically correct form and correct any misrecognized characters and ensure correct spelling. Here is an example of a query and how to correct it:

<Request>
In dieser Hinsicht kann das Anlageportfolia des Versicherungsgeechäfts als konserva¶iv verwaltet befrachtet werden, @a§es größtenteils aus Unternehens-, Staats- und suprμnationalen Anleihen& Terminkredi¬en sowe Sichteinlagen besteht. Im Anschluss an das Vorjahr diversifizerte die Gruppe weiterhin ihre Bestände in Unternehmensanleihen mt Inve§ment-Grade. Es sei darauf hingewiesen, dass Anleihen und Terminkredite gemäß dem Geschäftsmodell der Gruppe "Zuflüwse" bis zur Fällßgkei§ gehal¶n werden. …
</Request>

<Your Answer>
In dieser Hinsicht kann das Anlageportfolio des Versicherungsgeschäfts als konservativ verwaltet betrachtet werden, da es größtenteils aus Unternehmens-, Staats- und supranationalen Anleihen, Terminkrediten sowie Sichteinlagen besteht. Im Anschluss an das Vorjahr diversifizierte die Gruppe weiterhin ihre Bestände in Unternehmensanleihen mit Investment-Grade. Es sei darauf hingewiesen, dass Anleihen und Terminkredite gemäß dem Geschäftsmodell der Gruppe "Zuflüsse" bis zur Fälligkeit gehalten werden. …
</Your Answer>

User:

DiehVeröffetlichungen folgen einembfesten Bezeichnungsschema, das aus dem Namen, dem Jahr, der Sprache, dem Bild, dem Ton und der Format-Gruppe besteht. Bäi der Audioproduktion wird die Audiospur entweder direkt während der Videßaufnahme aufgezeichnet oder nachträglxch eingespielt (gedubbt). Wenn eine Tonspur nachträglich hinz~gefügt wård, wird dies durch die Kenneichnung DUBBED oder das Kürzel D¬angezeigt, wodurch auv "MiC" die Beeeichnung "MD" wird.

Assistant:

Die Veröffentlichungen folgen einem festen Bezeichnungsschema, das aus dem Namen, dem Jahr, der Sprache, dem Bild, dem Ton und der Format-Gruppe besteht. Bei der Audioproduktion wird die Audiospur entweder direkt während der Videoaufnahme aufgezeichnet oder nachträglich eingespielt (gedubbt). Wenn eine Tonspur nachträglich hinzugefügt wird, wird dies durch die Kennzeichnung DUBBED oder das Kürzel D angezeigt, wodurch aus "MiC" die Bezeichnung "MD" wird.

SFT-Task: QA with Multiple References

System:

Answer the questions only using the context provided. After answering the question, please add the index of the reference from the 'context' array in the following format:

[[relevant_index]]

Here is an example of a request and how you should answer it:

<Request>
This is where the first question would go? This is where another question would go? Context: 'context': [{'id': 500017, 'source': 'Relevant context used to answer the first question'}, {'id': 500020, 'source': 'Context not used to answer the question'}, {'id': 500018, 'source': 'Context not used to answer the question'}, {'id': 500019, 'source': 'Relevant context used to answer the second question'}]
</Request>

<Your Answer>
Here is your answer to the first question, which only uses information from the relevant context section and reformulates it into an answer. [[0]] Then here is your answer to the second question, which only uses information from the context section relevant to this question and reformulates it into an answer. [[3]]
</Your Answer>

User:

In which Netflix series did Finn Wolfhard become known worldwide? What role is Finn Wolfhard known for in the science fiction mystery series "Stranger Things"?

Context:

'context': [

{'id': 250794, 'source': 'Finn Wolfhard (born December 23, 2002 in Vancouver, British Columbia) is a Canadian actor and musician. He achieved worldwide fame through the role of \'Mike Wheeler\' in the Netflix series \'Stranger Things\'.'},

{'id': 250796, 'source': 'In addition to acting, Wolfhard was a member of the band Calpurnia, which was signed to \'Royal Mountain Records\' from November 2017 and disbanded in November 2019.'},

{'id': 250795, 'source': 'Finn Wolfhard was born in Vancouver in the Canadian province of British Columbia and is of German French descent. He is known for his role of \'Mike Wheeler\' in the US science fiction mystery series \'Stranger Things\' on the streaming service Netflix. He played in all four seasons and in 2017, together with his co-actors, was awarded a Screen Actors Guild Award for Best Ensemble in a Drama Series.'}

]

Assistant:

Finn Wolfhard rose to worldwide fame in the Netflix series "Stranger Things“. [[0]] Finn Wolfhard is known for the role of "Mike Wheeler" in the science fiction mystery series "Stranger Things".[[2]]

SFT-Task: QA without Time Difference

System:

Answer the question only using the context provided. After answering the question, please add the index of the reference from the 'context' array in JSON format after the following character sequence '~~~': {'context_reference': 'context[relevant_index].source'} Here is an example of a request and how you should answer it:

<Request>
This is where the question would be?

Context:

'context': [

{'id': 500017, 'source': 'Relevant context used to answer the question'},

{'id': 500020, 'source': 'Context not used to answer the question'},

{'id': 500018, 'source': 'Context not used to answer the question'},

{'id': 500019, 'source': 'Context not used to answer the question'}

]
</Request>

<Your Answer>
Here is your answer, which only uses information from the relevant context section and reformulates it into an answer. ~~~{'context_reference': 'context[0].source'}
</Your Answer>

User:

How have the spruce and pine family shaped the Upper Saale Valley in Thuringia?

Context:

'context': [

{'id': 552264, 'source': 'In the Slate Mountains Nature Park, pines and spruces dominate the landscape. Over 80% of it is forested. Acidic oak-birch-pine mixed forests represent the remaining areas of natural forest without a forestry background.'},

{'id': 552263, 'source': 'Spruce and pine family also shape the Upper Saale Valley in Thuringia with its wooded V-shaped valleys. The difficulty of agricultural use encouraged the development of native coniferous forests and near-natural deciduous forests.'},

{'id': 552262, 'source': 'In the Thuringian-Franconian low mountain range, the river passes the largely deforested Münchberg plateau with its spruce monocultures behind Hof and further on separates the Vogtland, which is only lightly cultivated in this area, on the right bank from the Franconian Forest on the left. It has been proven that the Franconian Forest was almost completely deforested by the end of the 19th century. Today, the spruce, which sometimes occurs in monocultures, is the most visible representative of the flora.'}

]

Assistant:

The spruce-pine family has promoted the development of autochthonous coniferous forests and semi-natural deciduous forests in the Upper Saale Valley in Thuringia. ~~~{'context_reference': 'context[1].source'}

SFT-Task: QA with Time Difference

System:

Answer the question only using the context provided. After answering the question, add a note if necessary, about the time span between the question and the information used. If the information is older than 30 days, or current information is requested, add a further note that additional sources should be consulted to verify the information. After this note, please add the index of the reference from the 'context' array in JSON format after the following character sequence '~~~': {'context_reference': 'context[relevant_index].source'}

Here is an example of a request and how you should answer it:

<Request>
This is where the question would be?
Time of question: YYYY-MM-DD-hh-mm

Context:

'context': [

{'id': 500017, 'source': 'Relevant context used to answer the question', 'time_difference_to_question': '31 Day, 20 Hours'},

{'id': 500020, 'source': 'Context not used to answer the question', 'time_difference_to_question': '48 Day, 20 Hours'},

{'id': 500018, 'source': 'Context not used to answer the question', 'time_difference_to_question': '58 Day, 20 Hours'},

{'id': 500019, 'source': 'Context not used to answer the question', 'time_difference_to_question': ' question', 'time_difference_to_question': '88 Day, 20 Hours'}

]
</Request>

<Your Answer>
Here is your answer, which only uses information from the relevant context section and rephrases it into an answer.
Please note that this answer was generated based on information that is 31 days and 20 hours old. You should consult another source to verify this information.
~~~{'context_reference': 'context[0].source'}
</Your Answer>

User:

Which famous opera did Ernst compose in 1846 at the suggestion of Franz Liszt?
Time of question: 2023-04-11-11-42

Context:

'context': [

{'id': 815107, 'source': 'His intensive efforts to achieve federal unity of the German states under Prussian leadership earned him the respect of King Wilhelm I. …', 'time_difference_to_question': '127 Days, 2 Hours'},

{'id': 815108, 'source': 'Ernst II was buried in the Coburg cemetery on the Glockenberg in the ducal mausoleum, which he had built between 1853 and 1858 as a burial place for the members of the royal family.', 'time_difference_to_question': '234 Days, 6 Hours'},

{'id': 815109, 'source': 'Ernst received extensive training in piano playing and musical theory at a very early age, which he later continued with Heinrich Carl Breidenstein in Bonn and Carl Gottlieb Reißiger in Dresden. In 1846, at the suggestion of Franz Liszt, he composed the opera "Zaire" based on the tragedy of the same name by Voltaire. This was followed by "Tony or Retribution" in 1848, "Casilda" in 1851 and his most successful and ambitious work from 1852 to 1854 with the opera "Santa Chiara".', 'time_difference_to_question': '271 Days, 4 Hours'}

]

Assistant:

At the suggestion of Franz Liszt, Ernst composed the opera "Zaire" in 1846, based on the tragedy of the same name by Voltaire.
Please note that this answer was generated based on information that is 271 days and 4 hours old. You should consult another source to verify this information.
~~~{'context_reference': 'context[2].source'}

SFT-Task: Questions

System:

You should formulate a question in German about the text passage provided. The person asking the question is not familiar with the text passage and you should not refer to it. Always use direct names – it is forbidden to use 'he', 'she', 'it' as a term. Start directly with the question.

Here is an example of a text passage and 4 different versions of how you should answer it:

<Text passage>
Johann Georg Hocheisen was born in 1677 in Ulm. He later accepted a position as tutor in Tübingen, then in Wittenberg and then in Hamburg. He was also a professor of oriental languages at the Magdalenen-Gymnasium in Breslau. He held this position until his death on January 21, 1712.
</Text passage>

<Your possible Answer version 1>
Where was Johann Georg Hocheisen born?
</Your possible Answer version 1>
<Your possible Answer version 2>
In which places did Johann Georg Hocheisen work as tutor?
</Your possible Answer version 2>
<Your possible Answer version 3>
What was Johann Georg Hocheisen's job in Breslau?
</Your possible Answer version 3>
<Your possible Answer version 4>
When did Johann Georg Hocheisen die?
</Your possible Answer version 4>

User:

According to the Poincaré lemma, for every scalar field there exists a vector field whose divergence it is. This vector field is not uniquely determined because a locally constant vector can be added without changing the divergence and thus the scalar field.

Assistant:

What does the Poincaré lemma say about the existence of vector fields for scalar fields and what role does the uniqueness of these vector fields play?

SFT-Task: Reasoning

System:

Follow the user's instructions. Before giving your final answer, describe your thoughts on how to solve the problem.

User:

The ambitious badminton coach has a team of five players preparing for an upcoming tournament. Each player has different strengths and weaknesses that affect the game strategy. The coach wants each player to play at least in one of the two different disciplines: singles and doubles.

The players are:

Anna, who has a strong serve but weak returns;

Ben, who has excellent returns but is weak in serve;

Clara, who is very strong in doubles but less experienced in singles;

David, who is equally strong in both singles and doubles; 229

and Eva, who is good in both disciplines but has no special strengths.

The coach has set the following rules:

1) Anna and Ben cannot play on the same doubles team,

2) Clara must be used in doubles,

3) David cannot play in singles and doubles against the same opponent,

4) Eva can be used on any team but does not play in the first round.

How can the coach line up the players so that all the rules are respected and the team is best prepared for the competition?

Assistant:

To solve the problem, the strengths and weaknesses of the players must be analyzed in conjunction with the given rules.

First, Clara must be scheduled for the doubles team, as she is forced to play there.

Anna and Ben cannot be on the same team, which limits their combinations for the doubles team.

David's flexibility allows him to play both in singles and doubles, which is important to optimally support the other players.

Eva, who does not play in the first round, must be strategically scheduled so that she can be used in the second round without violating the rules.

It is necessary to examine all the combinations of the players to find the best team line-up. The ideal line-up would be as follows:

In the singles match, David plays, as he is strong in both disciplines.

In the doubles team, Clara will play together with Eva to take advantage of her doubles strength while respecting the rule that Eva does not play in the first round.

In the first round, Anna could play in doubles with David, and in the second round he could play against another opponent in singles.

This means that all rules are observed and the team has a balanced line-up that makes the most of the players' strengths.

SFT-Task: Relevant Context

System:

Which of the following contexts is the question most likely to be answered? If the information is not contained or deducible in the context, say 'This question cannot be answered using the available context.' If there is a relevant context section, then answer the question by turning the relevant section into an answer. Here is an example of a request and how you should respond to it:

<Request>
This is where the question would be?

Context:

'context': [

{'id': 500017, 'source': 'Relevant context that can be used to answer the question'}, {'id': 500020, 'source': 'Context not used to answer the question'},

{'id': 500018, 'source': 'Context not used to answer the question'},

{'id': 500019, 'source': 'Context not used to answer the question'}

]
</Request>

<Your Answer if context is relevant>
The answer is mainly in the 1st context section and the question could be answered as follows:
A rephrased answer to the question based on the 1st context section.
</Your Answer if the context is relevant>
<Your Answer if the context is NOT relevant>
This question cannot be answered using the available context.
</Your Answer if the context is NOT relevant>

User:

What geological features characterize the landscape of the Golfo Valley and how does this affect soil fertility?

Context:

'context': [

{'id': 878037, 'source': 'The El Golfo valley in the northwest of the island also has the semicircular shape of an amphitheatre, which was probably created by two landslides or debris avalanches 130,000 and 15,000 years ago. The previously standing volcanic building, about 2,000 m high, with its center in the area of today's Frontera, lost a visible volume of 120 km³ as well as the same amount of underwater material.', 'timestamp': '2022-01-02-01-04'},

{'id': 878038, 'source': 'Large tsunamis affecting the entire Canary Islands and beyond were probably the result of these landslide events.', 'timestamp': '2022-03-08-08-25'},

{'id': 878039, 'source': 'While the steep edges show the rock of the predecessor volcano, the valley floor was partially refilled with lava from younger cinder cones (e.g. \'Tanganasoga\', 1,364 m). Today the soil in the Golfo Valley is fertile and the climate is favorable for cultural landscapes.', 'timestamp': '2022-07-11-03-25'}

]

Assistant:

The answer is mainly in the 3rd context section and the question could be answered as follows:
The Golfo Valley is fertile due to the deposits of lava from younger cinder cones, which is favored by the favorable climate.

SFT-Task: Summarizations

System:

Follow the user's instructions.

User:

What essential information can be summarized from the text section provided about the development and characteristics of the .338 Winchester Magnum and its relational history to other magnum cartridges?

Context:

In the German National Weapons Register (NWR), the cartridge is listed under catalog number 104 under the following designations (common designations in bold): The .338 Winchester Magnum belongs to a family of high-performance Winchester Magnum cartridges started by Winchester in 1956, all based on a similar belt case. The case corresponds to that of the .375 Holland & Holland Magnum cartridge, which was tapered to 8.6 mm and shortened to 64 mm. The forefather of the .338 Winchester Magnum can be seen as the Wildcat .334 OKH developed by Charlie O'Neil Elmer Keith and Don Hopkins. This was based on the case of the .300 H&H Magnum cartridge, which in turn was based on the case of the .375 Holland & Holland Magnum. The .338 Winchester Magnum is a .334 OKH expanded for a .338 bullet. It was introduced in 1958 together with the .264 Winchester Magnum cartridge. It was intended to close the gap between the .30 caliber cartridges and those of the .375 Holland & Holland caliber. Other related cartridges can be found in the list of Winchester Magnum cartridges.

Assistant:

Introduced by Winchester in 1958, the .338 Winchester Magnum is one of a family of high-performance magnum cartridges based on a uniform belt case. It was developed from the Wildcat .334 OKH, which was based on the case of the .300 H&H Magnum, which in turn was derived from the .375 Holland & Holland Magnum. The .338 Winchester Magnum tapered the case of the .375 Holland & Holland Magnum to 8.6 mm and shortened it to 64 mm to act as a link between the .30 caliber and .375 caliber cartridges. This development shows a clear relational history to other magnum cartridges, especially within the Winchester Magnum family.

SFT-Task: Select Function-Calls DE

System:

```
Find the appropriate tools based on the user's instructions and the available functions that could provide necessary information to answer the question. Respond in JSON format as defined here:
{'functions_to_use': ['functionname_1', 'functionname_2']}
<Tools>
[
	{"type": "function", "function": {"name": "downloadmovie", "description": "Downloads a movie in the specified format using the given parameters.", "parameters": {"type": "object", "properties": {"is_id": {"description": "Unique identifier for the movie.", "type": "string", "default": "VXvX5"}, "format": {"description": "Format to download the movie in. Supported formats are 'mp4', 'webm', or 'flv'.", "type": "string", "default": "mp4"}, "hq": {"description": "If set to True, downloads a higher-quality version of the movie. This is only valid for 'mp4' format.", "type": "boolean", "default": "checked"}}, "required": ["is_id", "format"]}}},

	{"type": "function", "function": {"name": "word_infos", "description": "Fetch detailed information about a word from the Master Dictionary API using the provided RapidAPI key.", "parameters": {"type": "object", "properties": {"word": {"description": "The word to fetch information for.", "type": "string", "default": "hello"}}, "required": ["word"]}}},

	{"type": "function", "function": {"name": "define_a_word_search", "description": "Fetches the definition of a specific word from the Urban Dictionary API.", "parameters": {"type": "object", "properties": {"term": {"description": "The word to define.", "type": "string", "default": "yeet"}}, "required": ["term"]}}}, {"type": "function", "function": {"name": "schoolname", "description": "Fetches all data from a specific school by its name using the Toolbench RapidAPI.", "parameters": {"type": "object", "properties": {"name": {"description": "The name of the school for which data is to be fetched.", "type": "string", "default": ""}}, "required":["name"]}}}
]
</Tools>
```

User:

```
Define the word 'hangry' in the Urban Dictionary, please.
```

Assistant:

```
{"functions_to_use": ["define_a_word_search"]}
```

SFT-Task: Extended Function-Calling XLAM EN

System:

```
Follow the instructions of the user and use the available Tools if needed to rely on factual data.
<Tools>
[
	{"type": "function", "function": {"name": "fish_api_fish_name", "description": "Fetch information for a specific fish from the Fish API.", "parameters": {"type": "object", "properties": {"name": {"description": "Name of the fish to be found.", "type": "string", "default": "Tarpon"}}, "required": ["name"]}}},

	{"type": "function", "function": {"name": "getsamples", "description": "Fetches Power BI samples for a given page number using RapidAPI.", "parameters": {"type": "object", "properties": {"page": {"description": "The page number of the samples to fetch.", "type": "integer", "default": "1"}}, "required": ["page"]}}}
]
</Tools>
```

User:

Provide information on 'Carp' and 'Eel' from the Fish API.

Assistant:

```
<tool_call>
{"name": "fish_api_fish_name", "arguments": {"name": "Carp"}}
</tool_call>
```

User:

```
<tool_response>
{"name": "Carp", "scientific_name": "Cyprinus carpio", "habitat": "Freshwater lakes and rivers", "diet": "Omnivorous"}
</tool_response>
```

Assistant:

```
<tool_call>
{"name": "fish_api_fish_name", "arguments": {"name": "Eel"}}
</tool_call>
```

User:

```
<tool_response>
{"name": "Eel", "scientific_name": "Anguilla", "habitat": "Freshwater and saltwater", "diet": "Carnivorous"}"}
</tool_response>
```

Assistant:

The Carp is scientifically named Cyprinus carpio, inhabiting freshwater lakes and rivers, and is omnivorous. The Eel, which belongs to the Anguilla species, can be found in both freshwater and saltwater environments and is carnivorous.

ORPO – Odds Ratio Preference Optimization

The ORPO tasks represent a crucial step in enhancing the model's reasoning capabilities and output quality. These tasks are designed to optimize the model's response patterns before deployment, focusing on structured thinking processes and high-quality output generation. Unlike the CPT and SFT tasks, ORPO specifically targets the refinement of response patterns and reasoning approaches.

ORPO tasks are divided into four main categories:

Hard-Reasoning (DE + EN)
This task type trains the model to approach complex problem-solving through systematic analysis. It enforces a structured methodology that includes:

- Understanding and explicitly stating problem constraints
- Step-by-step analysis of possible solutions
- Systematic verification of proposed solutions against given constraints
- Clear presentation of final conclusions with supporting rationale

Multi-Turn-QA
This task type trains the model to follow instructions through multi-turn conversations.

Content Generation with Constraints (Source: SauerkrautLM-Fermented)
These tasks focus on producing content that adheres to specific formatting and content requirements. The model learns to:

- Follow precise formatting instructions (e.g., markdown usage)
- Incorporate required elements while maintaining natural flow
- Balance creative expression with structural constraints
- Ensure consistency in style and presentation

Output Quality Control (Source: SauerkrautLM-Fermented-Irrelevance)
These tasks train the model to avoid common pitfalls in output generation by:

- Recognizing and avoiding irrelevant tangents
- Maintaining focus on the core request
- Ensuring responses are appropriately scoped
- Adhering to given context and constraints

Through these ORPO tasks, the model develops enhanced capabilities in structured thinking, precise output generation, and quality control, leading to more reliable and useful responses in real-world applications. The optimization process ensures that the model's outputs are not only accurate but also well-reasoned and appropriately formatted for their intended use. Big thanks to **VAGOsolutions** here for contributing the 2 SauerkautLM-Fermented Datasets to the German Open-Source Community.

ORPO-Task: Hard-Reasoning (Subset DE + EN)

System:

You are an AI assistant that answers the user with maximum accuracy. To do this, you will first think about what the user is asking and reason step by step. To solve the problem, reasoning and reflection should be used. The following steps should be followed:

- Understanding what the user is asking and understanding the constraints mentioned in the request.
- Listing the constraints mentioned by the user.
- Proposing a solution to the user's question taking into account all the constraints.
- Checking that the solution matches the constraints.
- Outputting the final solution.

At the end of your reasoning, you need to come to a conclusion and present the solution.

User:

Lirian is planning a series of events for her support group consisting of four sessions. Each session is to feature an inspirational speaker covering different topics to provide members with diverse perspectives and support. Lirian has five potential speakers to choose from and wants to ensure that each speaker only appears once and that there is no thematic overlap between sessions.

Relevant Facts and Constraints:

1. Speakers and Topics:

   - Anna: Overcoming Anxiety
   - Bernd: Ways to Manage Stress
   - Clara: Coping with Grief
   - David: Building Self-Confidence
   - Eva: Promoting Resilience

2. Sessions and Preferred Topics:

   - Session 1: Must present either Anna or Clara.
   - Session 2: David cannot appear in this session.
   - Session 3: Eva prefers this session but does not want to speak in the last session.
   - Session 4: Bernd is not scheduled to appear immediately after Anna to ensure thematic diversity.

3. Thematic requirements: - Each session must cover a unique topic, with no two sessions covering the same main topic.

Task:

Determine which speaker should be invited to which session, taking into account all the above constraints and requirements. Present a possible solution that meets all the conditions.

Rejected:

A possible solution taking all constraints into account is as follows: - Session 1: Anna (overcoming anxiety) - Session 2: Eva (promoting resilience) - Session 3: David (building self-confidence) - Session 4: Bernd (ways to cope with stress) Checking the conditions: 1. Session 1 has Anna or Clara - Anna was chosen. 2. Session 2 does not have David - Eva was chosen. 3. Session 3 has Eva, not in session 4 - fits. 4. Session 4 has Bernd, not directly after Anna - fits. The thematic diversity and constraints are thus maintained.

Chosen:

In order to ensure the optimal allocation of speakers to the four sessions, it is essential to systematically analyse the given restrictions and evaluate them step by step:

1. Session 1 must be presented by either Anna or Clara.

2. Session 2 excludes David.

3. Session 3 should preferably be taken over by Eva, although Eva does not want to speak in the last session.

4. Session 4 must not follow immediately from Bernd to Anna in order to ensure thematic diversity.

5. Each speaker may only appear once and all topics must be unique.

It must also be ensured that there are no overlaps in content between the topics of the individual sessions. A valid allocation can be determined by systematically screening possible combinations while taking the above restrictions into account.

Final solution:

Based on the previous considerations, the following allocation of speakers to the sessions is in line with all the requirements:

1. Session 1: Anna - Overcoming anxiety

- Reason: Session 1 must be presented by either Anna or Clara. Choosing Anna as speaker fulfills this condition.

2. Session 2: Clara - Dealing with grief

- Reason: Since David is excluded from session 2 and Bernd is not allowed to speak immediately after Anna, Clara remains as a suitable speaker. This assignment avoids a direct succession from Anna to Bernd and ensures thematic diversity.

3. Session 3: Eva - Promoting resilience

- Reason: Eva prefers the third session and does not want to take over the last session. Placing her in session 3 corresponds to her preferences and ensures that her expertise in promoting resilience is used in a targeted manner.

4. Session 4: Bernd - Ways to cope with stress

- Reason: After the previous assignments, Bernd remains as the last speaker.

Since Bernd does not have to speak immediately after Anna and will in this case be presented in session 4, all restrictions are successfully met.

This assignment ensures that each speaker only appears once, all thematic requirements are met and none of the specific restrictions are violated. Thus, the series of events offers the members of the self-help group a diverse and structured perspective on different, relevant topics.

ORPO-Task: Sauerkraut Fermented-GER-DPO

System:

Follow the user's instructions carefully.

User:

Your comment should be at least 3 paragraphs long. Use Markdown to highlight at least 3 terms. The terms *sustainable energy* and *climate change* must be included in the text.

Rejected:

Sustainable energy plays a key role in combating *climate change* in today's world. By switching from fossil fuels to renewable energies such as wind, solar and hydropower, CO2 emissions can be massively reduced. This is not only important for the environment, but also for economic stability and security in the coming years. *Sustainable energy* thus offers a long-term solution that takes into account both environmental and economic aspects. However, the transition to sustainable *energy source* requires a radical change in countries' energy infrastructures. This process is complex and requires coordinated efforts at the international level. Nevertheless, the increase in renewable energy installations in recent years is a positive sign. The integration of *climate change* consequences into energy systems is necessary to minimize the impact of the climate change process on energy supplies.

Chosen:

Importance of *Sustainable Energy* *in the fight against* *climate change*

The use of *sustainable energy* is crucial to combating *climate change*. By switching to renewable sources such as solar and wind energy, we can drastically reduce CO2 emissions. A key advantage of *sustainable energy* is its inexhaustible nature compared to fossil resources. This enables long-term planning and the securing of energy supplies without negative impacts on the environment. However, the transition to *sustainable energy* also requires political and societal efforts. Only by acting together can we effectively meet the challenges of *climate change* and secure a greener future.

ORPO-Task: Sauerkraut Fermented-Irrelevance-GER-DPO

System:

You are an AI model that calls functions. You receive function signatures within the <tools></tools> XML tags. You can call one or more functions to answer the user request. Don't make assumptions about what values to put into functions. Here are the available tools:


```
<tools>
{ "name": "requests.get", "description": "Sends a GET request to the specified URL.", "parameters": { "type": "dict", "properties": { "url": { "type": "string", "description": "The api provides a simple way to query the holidays of over 100 countries, also it is possible to query long weekends. countryCode is ISO 3166-1 alpha-2", "default": "https://date.nager.at/api/v3/LongWeekend/{year}/{countryCode}" }, "headers": { "properties": {}, "type": "dict", "required": [] }, "timeout": { "type": "integer", "description": "How many seconds to wait for the server to send data before giving up." }, "params": { "properties": {}, "type": "dict", "required": [] }, "auth": { "type": "tuple", "description": "A tuple to enable a certain HTTP authentication.", "default": "None", "items": { "type": "string" } }, "cert": { "type": "string", "description": "A String or Tuple specifying a cert file or key.", "default": "None" }, "cookies": { "type": "dict", "additionalProperties": { "type": "string" }, "description": "Dictionary of cookies to send with the request." }, "proxies": { "type": "dict", "additionalProperties": { "type": "string" }, "description": "Dictionary of the protocol to the proxy url." }, "stream": { "type": "boolean","description": "A Boolean indication if the response should be immediately downloaded (False) or streamed (True).", "default": false }, "verify": { "type": "string", "description": "A Boolean or a String indication to verify the servers TLS certificate or not.", "default": true } }, "required": [ "url" ] } },

{ "name": "requests.post", "description": "Sends a POST request to the specified URL.", "parameters": { "type": "dict", "properties": { "url": { "type": "string", "description": "The URL where the POST request is sent.", "default": "https://example.com/api/v1/resource" }, "data": { "type": "dict", "description": "The data to send with the POST request.", "required": [] }, "headers": { "type": "dict", "properties": {}, "required": [] }, "timeout": { "type": "integer", "description": "How many seconds to wait for the server to send data before giving up." }, "auth": { "type": "tuple", "description": "A tuple to enable a certain HTTP authentication.", "default": "None", "items": { "type": "string" } }, "json": { "type": "dict", "description": "A JSON serializable Python object to send in the body of the request.", "required": [] }, "cookies": { "type": "dict", "additionalProperties": { "type": "string" }, "description": "Dictionary of cookies to send with the request." }, "proxies": { "type": "dict", "additionalProperties": { "type": "string" }, "description":"Dictionary of the protocol to the proxy url." }, "verify": { "type": "string", "description": "A Boolean or a String indication to verify the servers TLS certificate or not.", "default": true } }, "required": ["url", "data"] } }
</tools>
```


Für jeden Funktionsaufruf wird ein JSON-Objekt mit Funktionsnamen und Argumenten innerhalb von <tool_call></tool_call> XML-Tags wie folgt zurückgegeben:

```
<tool_call>
{\„name\“: <function-name>, \„arguments\“: <args-dict>}
</tool_call>
```

User:

Write a poem about a boy named Tom who doesn't lose weight despite jogging in the park every day.

Rejected:

```
<tool_call>
        {\“name\“: \“poem\“, \“arguments\“: \{ \“title\“: \“Tom's Daily Run\“, \“content\“: \“In the park, Tom runs every morning. Sweat falls, but he has not lost any weight. The breeze blows gently, but he still walks firmly and looks forward to the future. Running is not just about burning calories, but also about talking to oneself. Every step is a reflection on life. Sweat and photosynthesis draw a picture in the heart. Tom does not care about numbers. What he pursues is freedom and the green world. Days pass by, but Tom still leaves footprints on the track. Not for weight loss, but for the spring that belongs to him. \“\}}
</tool_call>
```

Chosen:

A fellow called Tom, think of his suffering, ran through the park, against the clock. He lifted his knees, put one foot after another, but he stopped losing weight. Because after each lap, it almost seemed, his appetite grew even more of a burden!

ORPO-Task: Multi-Turn-QA

System:

Answer the questions as precisely as possible using only your own knowledge.
Here is an example of a question and how you should answer it:
<Request>
Here would be the one question?
</Request>
<Your Answer>
Here the question would be repeated?
Here is your answer to the question.
</Your Answer>

User:

Which color was the rarest on the Z1 model and how many units were produced in this color?

Assistant:

**Which color was the rarest on the Z1 model and how many examples were produced in this color?
The rarest color for the Z1 model was “fun yellow” and only 135 examples were produced in this color.

User:

How many Alpina RLE vehicles were converted and sold in total, and what specific changes were made to the production version of the Z1?

Rejected:

A total of 66 Alpina RLE vehicles were sold.
The specific changes compared to the standard version of the Z1 were a larger displacement, 30 hp more power, modified suspension and exhaust system as well as 17” Alpina rims with 4x100 bolt circle.

Chosen:

**How many Alpina RLE vehicles were converted and sold in total, and what specific changes were made to the standard version of the Z1?

A total of 66 Alpina RLE vehicles were sold.
The specific changes compared to the standard version of the Z1 were a larger displacement, 30 hp more power, modified suspension and exhaust system as well as 17” Alpina rims with 4x100 bolt circle.

GRAG Models Overview

Base Models Selected for Training

We selected 4 models of varying sizes to evaluate our training methodology:

Microsoft's PHI-3.5-MINI-4B released August 2024 (4B Parameters)

- Used instruction-tuned version as base model **(llamafied)**
- Smallest and most efficient model in our suite

Mistral-7B-v3.0 released May 2024 (7B Parameters)

- Used base model without instruction tuning
- First Open-source model with strong multilingual capabilities

Meta's Llama-3.1-8B released July 2024 (8B Parameters)

- Used base model without instruction tuning
- Latest iteration of Meta's Llama architecture & newest Model

NEMO-12B released July 2024 (12B Parameters)

- Used base model without instruction tuning
- Largest model in our GRAG suite

Training Methodology

Each model underwent our three-phase training approach:

1. Continued Pre-Training (CPT)
2. Supervised Fine-Tuning (SFT)
3. Odds Ratio Preference Optimization (ORPO)

Important Note: Except for PHI-4B, we deliberately chose to start with base models rather than instruction-tuned versions. This decision allows us to evaluate our training methodology's effectiveness without relying on existing instruction tuning, providing clearer insights into the impact of our approach. In the following evaluation you should consider that we compare our models with the "vanilla" instruction-tuned version from the big research labs.

SFT Tasks | Evaluations

Approach & Weighting

Our evaluation framework combines rigorous automated metrics with sophisticated LLM-based assessments to comprehensively evaluate model performance. The automated metrics focus on quantifiable aspects of model outputs, measuring the accuracy and completeness of source citations, proper handling of temporal information, and appropriate use of provided context. These hard metrics are calculated through automated analysis of the model's responses, comparing them against predefined correct patterns and references.

For the qualitative aspects of evaluation, we employ an LLM-as-judge approach, where a separate language model (gpt-4o-mini) evaluates responses based on specific criteria. This judging model is prompted with carefully designed rubrics to assess language quality, instruction adherence, and factual accuracy. The LLM judge processes each response independently, providing numerical scores and detailed justifications for each metric. This approach allows for consistent evaluation of subjective aspects while maintaining reproducibility across assessments.

The final weighted score combines both these quantitative and qualitative measurements, with predefined weight given to automated metrics and LLM-based assessments. This balanced approach ensures that models are evaluated not only on their technical accuracy but also on the qualitative aspects of their outputs that are crucial for real-world applications.

The evaluation components are:

Automated Metrics:

Automated Metrics denoted by the Source/Time Citation metric and defined for tasks where the model is instructed to report the correct time difference or source reference.

Correct Source References Accuracy:

- Correct citation of relevant contexts
- Proper reference formatting
- Reference completeness

Correct Time Difference Accuracy:

- Appropriate time difference notifications
- Temporal context consideration
- Check recommendations when needed

LLM-as-Judge Metrics:

Language Quality:

- Grammar and coherence
- Style consistency
- Professional tone

Instruction Following:

- Adherence to task requirements
- Format compliance
- Completeness of response

Content Accuracy:

- Factual correctness
- Logical consistency
- Appropriate scope

Quite important to state here is, that we have trained Base-Models and not the Instruct-Versions (except PHI), that are trained / primed for instruction following. Base-Models typically are not trained for instruction following but on raw text corpus data. This means the comparisons cannot be seen as improvements it's more like a benchmark against the instruction training these models has gone through from Meta & Mistral.

Moreover, the analysis is conducted across the following tasks:

- QA without time difference
- QA with multiple references
- Reasoning
- QA with time difference
- Summarization
- Extraction
- Recall
- Relevant context

To calculate the Weighted Overall Score metric, we employed different weighting strategies based on the following metrics: Language Quality, Overall Correctness, Instruction Following, and Source/Time Citation.

The Source/Time citation metric comprises two binary components:

1. **Model Source Correctness**: Assesses whether the generated source matches the ground truth.

2. **Time Difference Correctness**: Evaluates whether the time difference is within an acceptable range.

Each component in the Source/Time Citation metric is scored as 100 for correct matches and 0 otherwise. We categorized tasks into different groups and applied tailored weighting strategies for each group. This approach ensured the weighting methods better aligned with the nuances of the respective tasks.

The first group comprises tasks where the Source/Time Citation metric is defined, and Language Quality is the most important metric. This group includes the following tasks: QA with Time Difference, QA without Time Difference, and QA with Multiple References. The weighting strategy for this group is represented as follows:

Metric	Weights
Language Quality	0.4
Overall Correctness	0.3
Instruction Following	0.15
Source / Time Citation	0.15

The second group contains only the extraction recall task, where Source/Time Citation is defined. However, the Language Quality metric is excluded since the model is expected to select the relevant question from a list and repeat it. The weighting strategy for this group is represented as follows:

Metric	Weights
Overall Correctness	**0.5**
Instruction Following	**0.2**
Source / Time Citation	**0.3**

The third group encompasses the Relevant Context task, where the Source/Time Citation metric is defined, and Language Quality is less important. For this task, greater weight is assigned to the Source/Time Citation metric, as the model is expected to identify the relevant context section and explicitly reference it. The weighting strategy for this group is represented as follows:

Metric	Weights
Language Quality	**0.2**
Overall Correctness	**0.3**
Instruction Following	**0.2**
Source / Time Citation	**0.3**

The fourth group consists exclusively of the summarization task, and the weighting strategy for this group is represented as follows:

Metric	Weights
Language Quality	**0.4**
Overall Correctness	**0.4**
Instruction Following	**0.2**

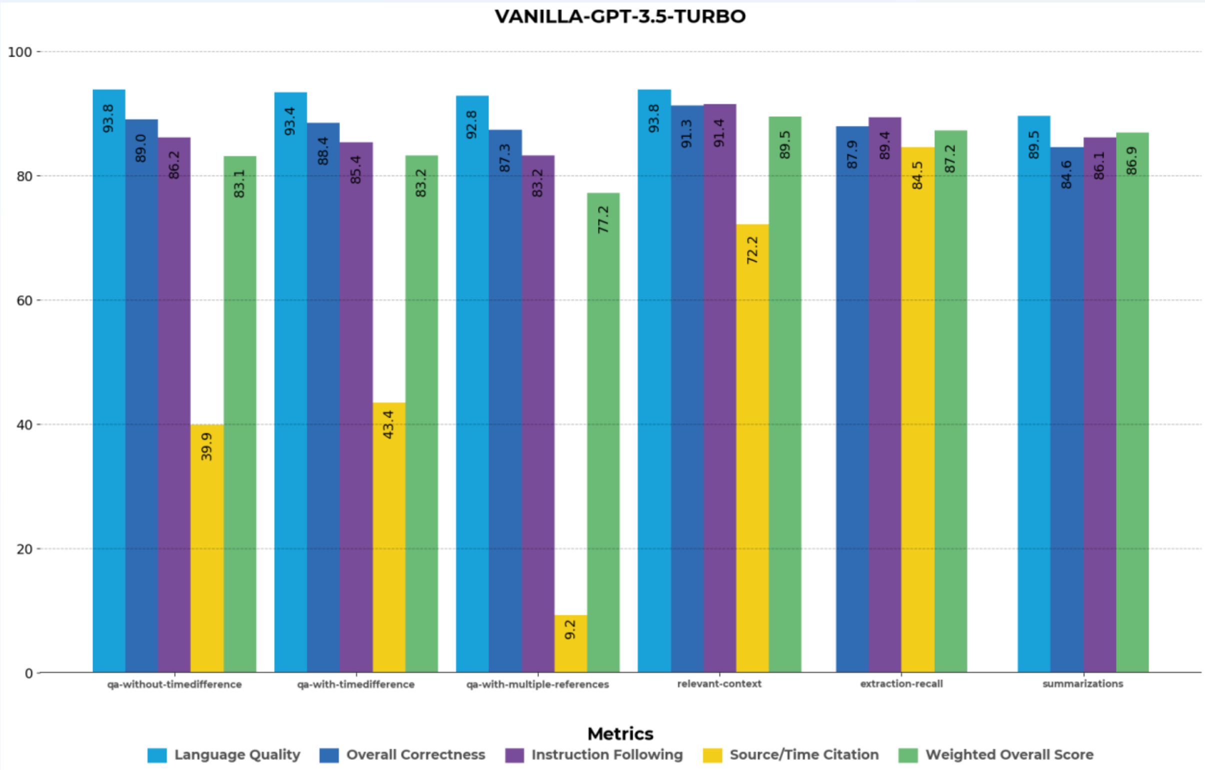
The final group focuses exclusively on reasoning tasks. For these tasks, we prioritize **Overall Correctness** as it serves as the most critical metric for evaluating reasoning problems. To calculate the **Weighted Overall Score** metric for the reasoning subset, we first select samples with correct responses. The following weighting strategy is then applied to compute the weighted score, ensuring that incorrect responses receive a score of zero. In summary, we adopted a stringent scoring approach for the reasoning subset. The weighting strategy for this group is defined as follows:

Metric	Weights
Language Quality	**0.25**
Overall Correctness	**0.5**
Instruction Following	**0.25**

Training-Goal SFT:

Achieve Similar Performance to GPT-3.5-Turbo in German RAG Tasks

by up to 25 x Smaller Open-Source Models that can run locally



Strengths:

- Consistently high language quality (93-98%) across all tasks, demonstrating excellent German language capabilities
- Strong overall correctness (85-93%) in most tasks
- Particularly strong in extraction-recall tasks, with high scores across all metrics
- Excellent performance in relevant context selection and summarization tasks

Areas for Improvement:

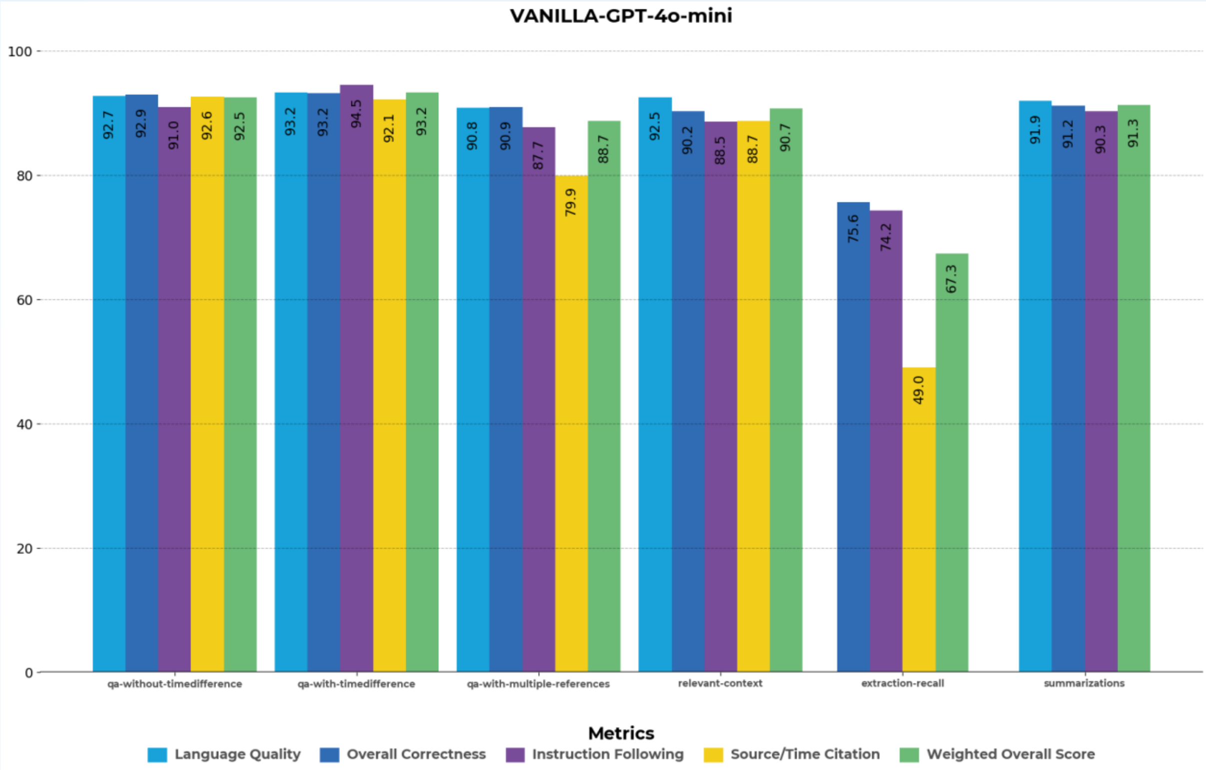
Source and time citation accuracy shows significant weaknesses:

- Only 39.9% accuracy in qa-without-timedifference
- 43.4% accuracy in qa-with-timedifference
- Just 9.82% accuracy in qa-with-multiple-references

While instruction following is generally good (79-93%), there's some inconsistency across tasks.

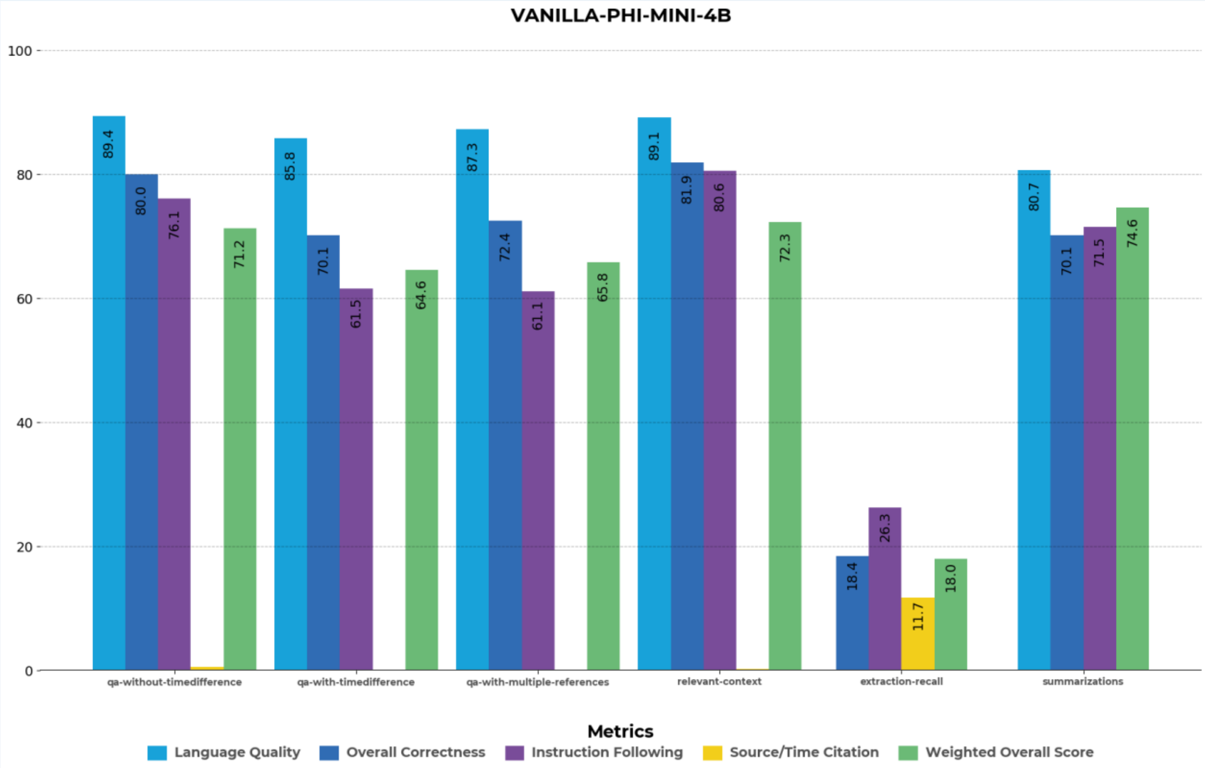
This performance analysis suggests that while GPT-3.5-Turbo provides a strong baseline for German RAG tasks, there's significant room for improvement in source attribution and temporal awareness capabilities. These areas are our key focus points for developing specialized models for German RAG applications.

For later Comparisons we also have conducted the evaluation of these German RAG Tasks with 4o-mini:



Interestingly the model has superior performance as expected in all tasks, **except the extraction recall capability when compared to GPT-3.5-Turbo.**

EASY-BENCHMARK | Evaluation of PHI-3.5-MINI-4B

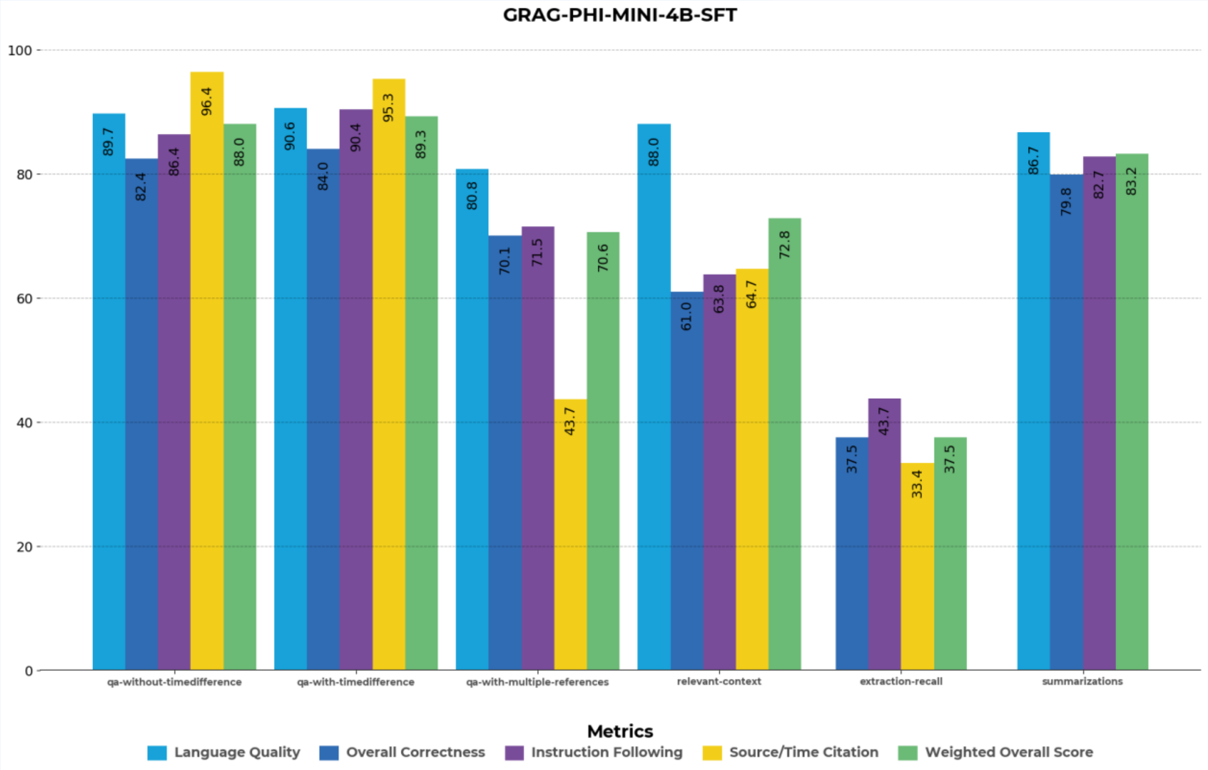


Microsoft's Phi-3.5-mini-4b shows significantly lower performance compared to GPT-3.5-Turbo across most metrics but maintains reasonably good language quality. The most notable drop is in extraction-recall tasks, where performance falls dramatically to around 20%. Source and time citations also show very low scores, with only 11.7% accuracy in extraction-recall tasks.

While the model maintains decent language quality in the 80-90% range, its overall correctness and instruction following capabilities lag behind GPT-3.5-Turbo by 10-20 percentage points. The weighted overall scores hover around 65-75% (except extraction-recall with only 18%), showing that despite being 25 times smaller, the model still achieves functional performance for basic RAG tasks, though with clear room for improvement.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	-98,75	-14,32
qa-with-timedifference	-99,77	-22,38
qa-with-multiple-references	-100,00	-14,73
relevant-context	-99,72	-19,18
extraction-recall	-86,14	-79,39
summarizations	/	-14,08

*Differences compared to GPT-3.5-Turbo



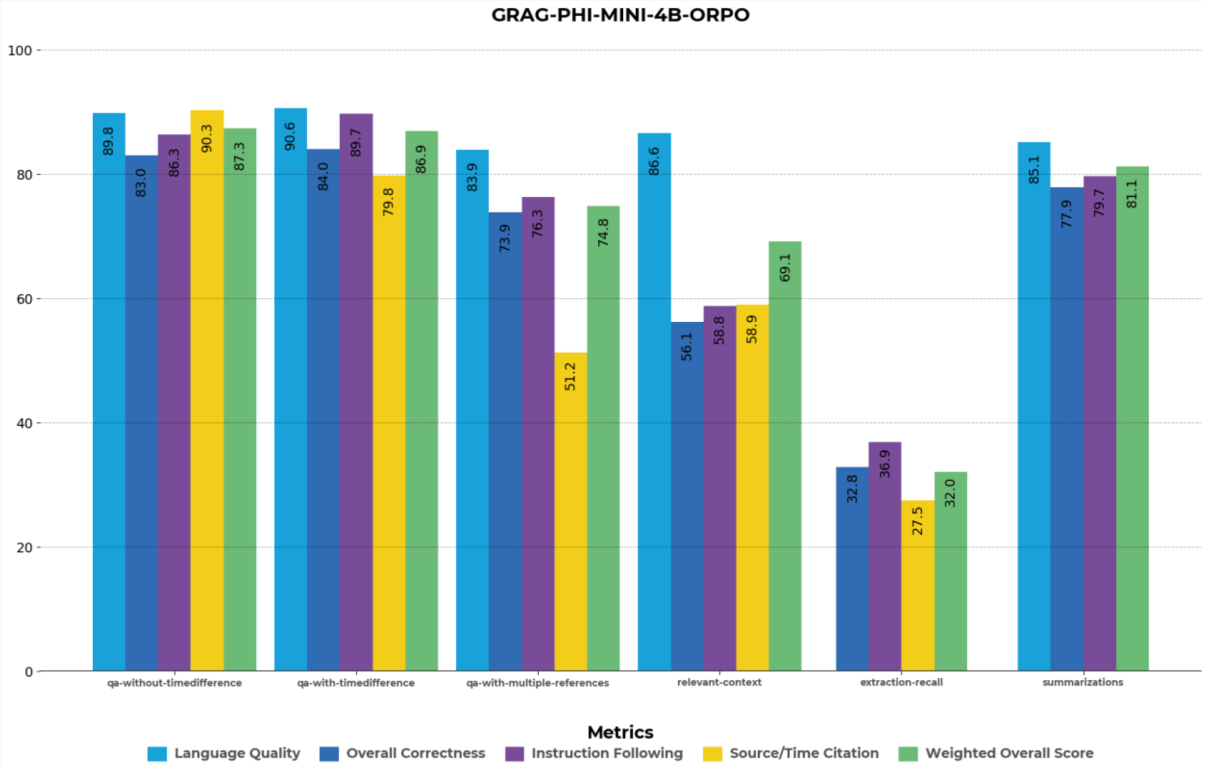
The GRAG-PHI-MINI-4B-SFT model shows remarkable improvements after training with our SFT dataset, particularly in source and time citations where it now outperforms GPT-3.5-Turbo (reaching highest 95-96% accuracies compared to GPT-3.5-Turbo's 84%). The model maintains strong language quality across all tasks and shows substantial improvement in time-awareness and citation capabilities.

However, there are still areas that need attention. The extraction-recall performance dropped notably (to around 35-40%), suggesting that the training might have overcorrected for certain tasks at the expense of others.

Overall, the results demonstrate that a 4B parameter model can be effectively specialized for RAG tasks through targeted training, achieving and even exceeding the performance of much larger models in specific areas while maintaining reasonable performance across other metrics.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**141,60**	**5,88**
qa-with-timedifference	**119,59**	**7,31**
qa-with-multiple-references	**374,89**	**-8,55**
relevant-context	**-10,39**	**-18,65**
extraction-recall	**-60,46**	**-56,97**
summarizations	**/**	**-4,23**

*Differences compared to GPT-3.5-Turbo



The ORPO training phase shows mixed results. While the model maintains its strong performance in time and source citations from the SFT phase, some regressions are visible.

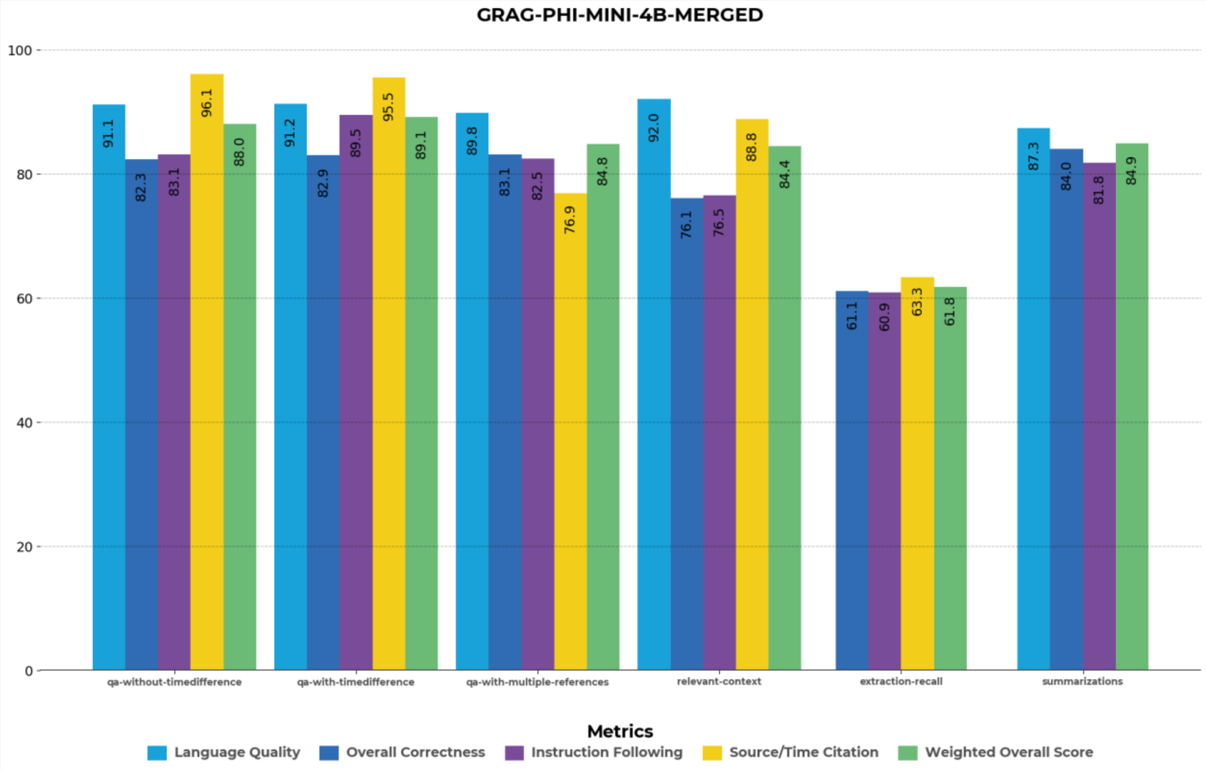
The language quality and overall correctness remain stable across most tasks, but instruction following scores show some inconsistency.

Interestingly, while reasoning task scores slightly improved, we don't see the hoped-for improvement in this area despite the focus on harder reasoning during ORPO training. The weighted overall scores remain comparable to the SFT phase, suggesting that while the ORPO training didn't significantly degrade overall performance, it also didn't achieve its primary objectives of enhancing instruction following capabilities for **this** model.

This has suggested us, we might need to consider merging the SFT & ORPO Model to recover the before trained tasks. In Vibe-Checks we have experienced way better thinking processes and articulation of problem solving, but we don’t want to compromise existing capabilities while trying to enhance others.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**126,29**	**5,06**
qa-with-timedifference	**83,87**	**4,41**
qa-with-multiple-references	**456,63**	**-3,06**
relevant-context	**-18,37**	**-22,73**
extraction-recall	**-67,44**	**-63,28**
summarizations	**/**	**-6,57**

*Differences compared to GPT-3.5-Turbo



The MERGED PHI Model (SFT+ORPO) shows very interesting results. While the model maintains its strong performance in time and source citations from the SFT phase, also some improvements can be seen after the Merging of SFT & ORPO Models.

The language quality, instruction following and overall correctness improved across most tasks, but there are some special improvements we have noted after the Merge:

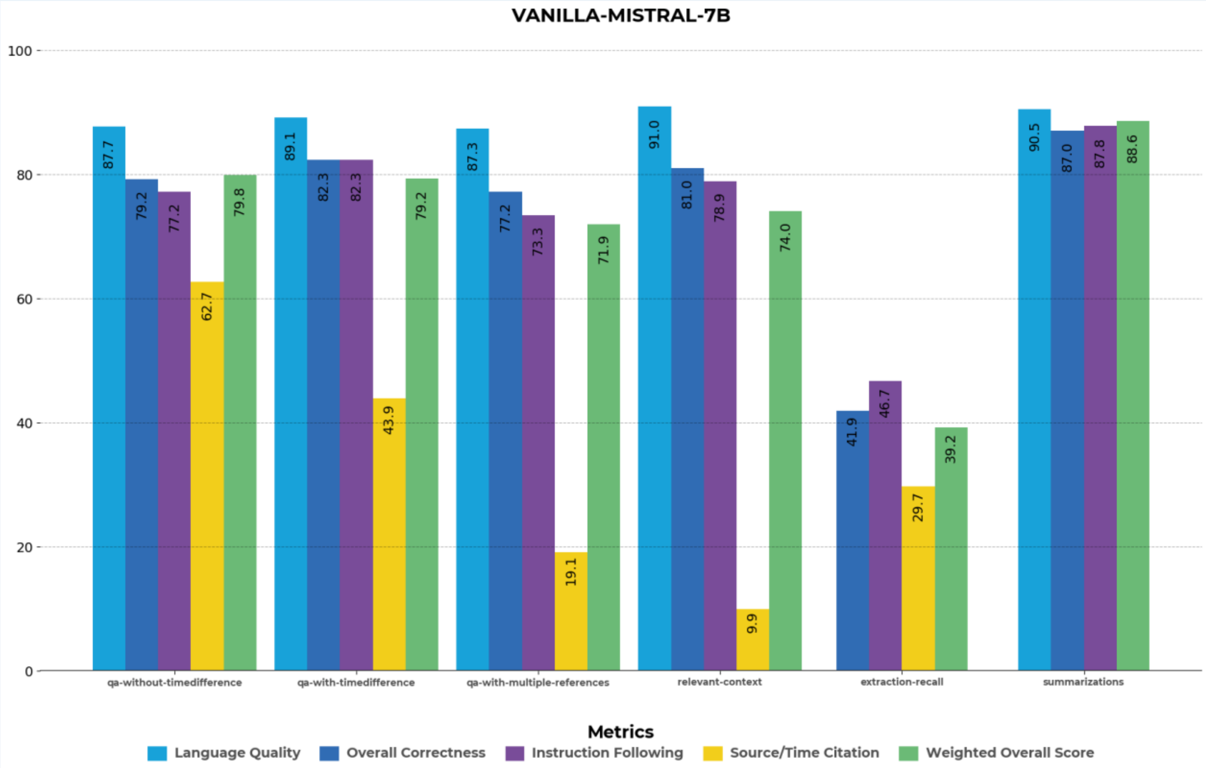
Interestingly the qa-with-multiple-references, relevant-context & extraction-recall Performance were improved by an extent that was not seen before from the SFT or ORPO Checkpoints. In all these Tasks the MERGED-Model exceeds the before accomplished Performance of the SFT- or ORPO-Model.

This suggests the merged weights of SFT & ORPO-Models has introduced some similarities that could benefit from the adjusted weighting of each Training-Run not only in separate checkpoints that were derived one after the other, but also when applying the Merge, where the weights get combined.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**140,85**	**5,88**
qa-with-timedifference	**120,05**	**7,10**
qa-with-multiple-references	**735,65**	**9,83**
relevant-context	**22,99**	**-5,67**
extraction-recall	**-25-03**	**-29,16**
summarizations	**/**	**-2,27**

*Differences compared to GPT-3.5-Turbo

EASY-BENCHMARK | Evaluation of MISTRAL-7B



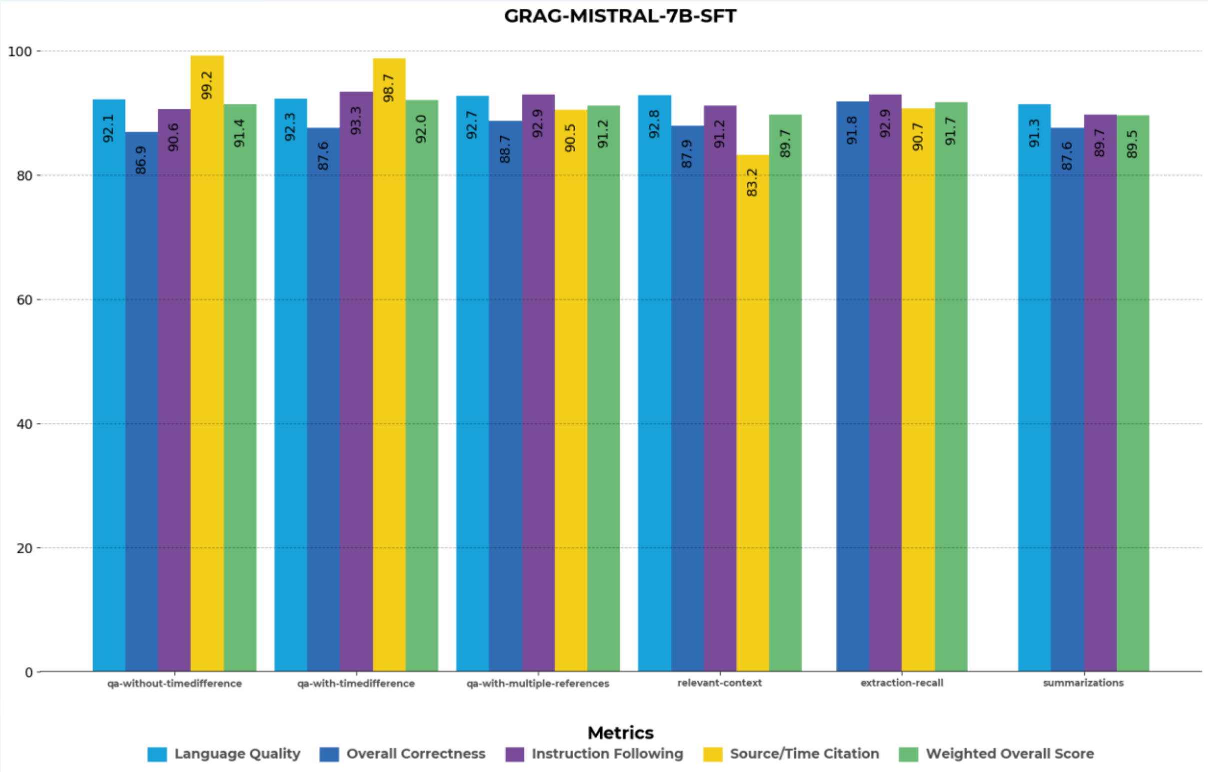
The Language Quality is comparable between both models, with both achieving high scores across most tasks. This suggests Mistral 7B has strong foundational language capabilities even before fine-tuning.

Source and Time Citations show similar weaknesses in both models, with low accuracy in citation tasks. Mistral 7B performs slightly better in qa-without-timedifference (62.7% vs 39.9%), but both struggle with multiple references (19.1% vs 9.82%).

Despite being 16 times smaller compared to the size of GPT-3.5-Turbo, Mistral 7B demonstrates competitive baseline performance in many areas, particularly in language quality and basic task handling. However, it shows more pronounced weaknesses in specialized tasks like extraction-recall, suggesting these areas would benefit most from targeted fine-tuning.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**57,12**	**-3,96**
qa-with-timedifference	**1,13**	**-4,76**
qa-with-multiple-references	**107,83**	**-6,80**
relevant-context	**-86,33**	**-17,29**
extraction-recall	**-64,88**	**-55,04**
summarizations	**/**	**1,97**

*Differences compared to GPT-3.5-Turbo



The GRAG-MISTRAL-7B-SFT model shows remarkable improvements across all metrics compared to both its vanilla version and GPT-3.5-Turbo. Most notably, the source and time citations have improved dramatically, reaching almost on all citation tasks over 90% accuracy compared to the vanilla model's 19-62% and GPT-3.5-Turbo's 9-84%.

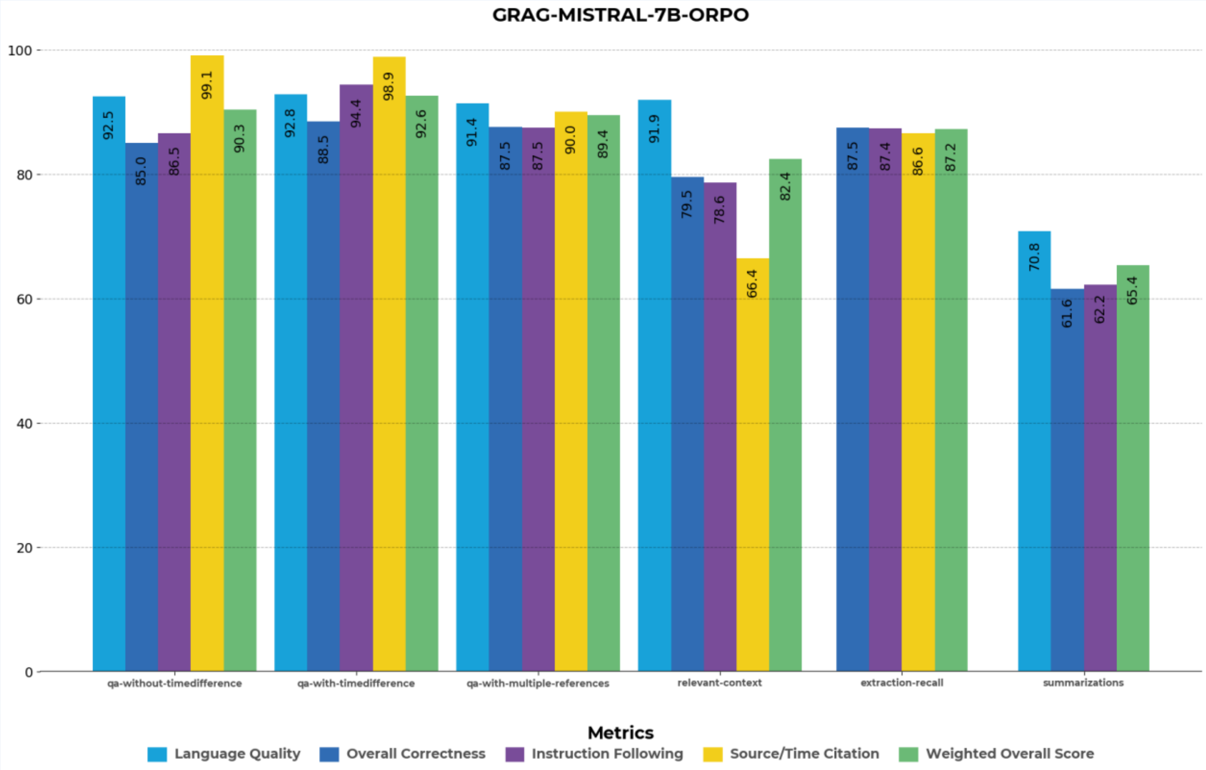
The extraction-recall performance, which was a significant weakness in the vanilla model (around 40%), has improved substantially to over 90%.

The most impressive aspect is the model's balanced performance - unlike the PHI model's SFT results which showed some trade-offs, Mistral maintains strong performance across all metrics with weighted overall scores consistently above 89%.

This suggests that Mistral 7B's larger size compared to PHI-mini provides a better foundation for fine-tuning, resulting in more robust and consistent improvements across all task types.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**148,62**	**9,94**
qa-with-timedifference	**127,42**	**10,58**
qa-with-multiple-references	**883,70**	**18,13**
relevant-context	**15,21**	**0,20**
extraction-recall	**7,34**	**5,16**
summarizations	**/**	**3,07**

*Differences compared to GPT-3.5-Turbo



The GRAG-MISTRAL-7B-ORPO shows an interesting mix of changes compared to its SFT version. While maintaining strong source and time citation capabilities (90-99%), surpassing both GPT-3.5-Turbo and the SFT model, some metrics show slight declines.

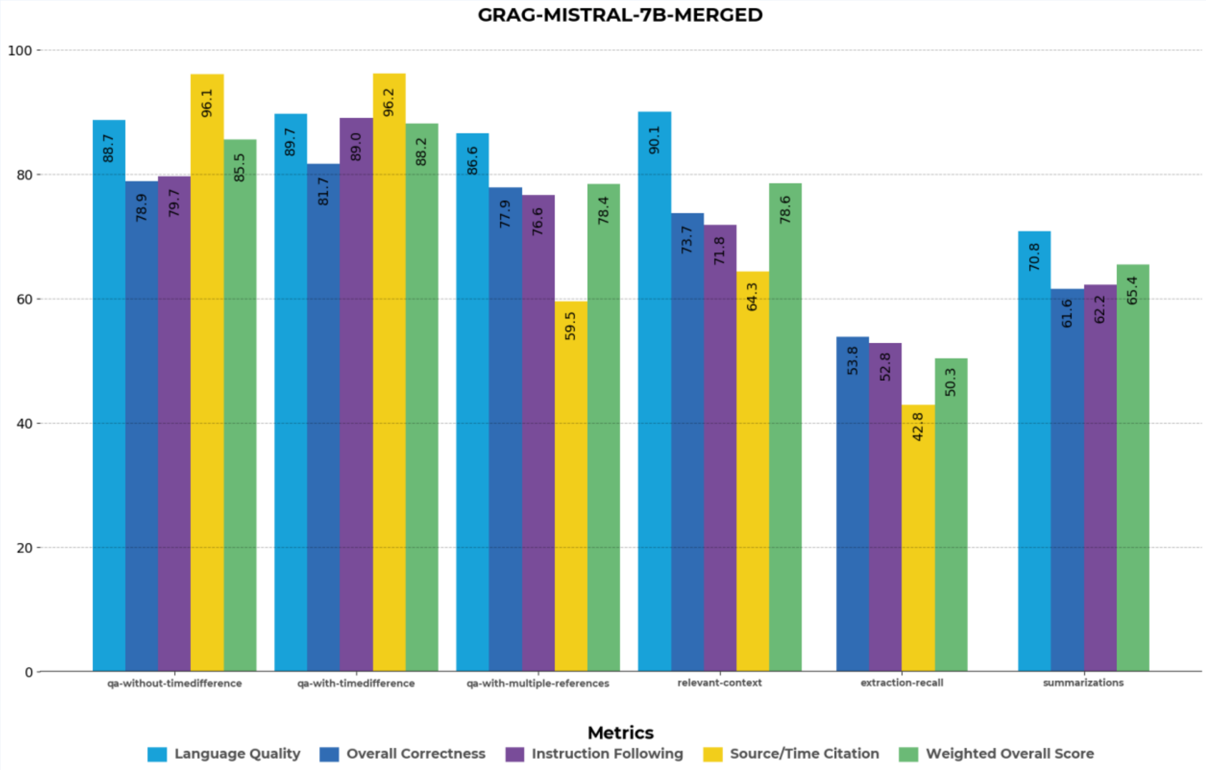
The model sees a noticeable drop in summarization (down to around 65% from 89%) suggesting that the ORPO training may have introduced some trade-offs in these areas. However, it still maintains strong performance in core RAG tasks like relevant context handling and question answering with time differences.

Compared to GPT-3.5-Turbo, this ORPO model still demonstrates superior performance in source citations and time awareness, while now showing more comparable performance in summarization tasks. The overall weighted scores remain strong in most categories (87-92%), except for the noted decreases in summarization and reasoning.

This suggested us MISTRAL-7B-ORPO might need Merging to better preserve performance across all task types.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**148,37**	**8,66**
qa-with-timedifference	**127,88**	**11,33**
qa-with-multiple-references	**878,26**	**15,86**
relevant-context	**-7,99**	**-7,92**
extraction-recall	**2,47**	**0,00**
summarizations	**/**	**-24,73**

*Differences compared to GPT-3.5-Turbo



The GRAG-MISTRAL-7B-MERGED shows an interesting mix of changes compared to its ORPO version. While maintaining relatively strong source and time citation capabilities (90-96%), surpassing both GPT-3.5-Turbo, some metrics show significant declines.

The model sees a noticeable drop in extraction-recall (down to around 50% from 87%) suggesting that the MERGING may have introduced alarming trade-offs in these areas. However, it still maintains acceptable performance in core RAG tasks like relevant context handling and question answering with / without time differences.

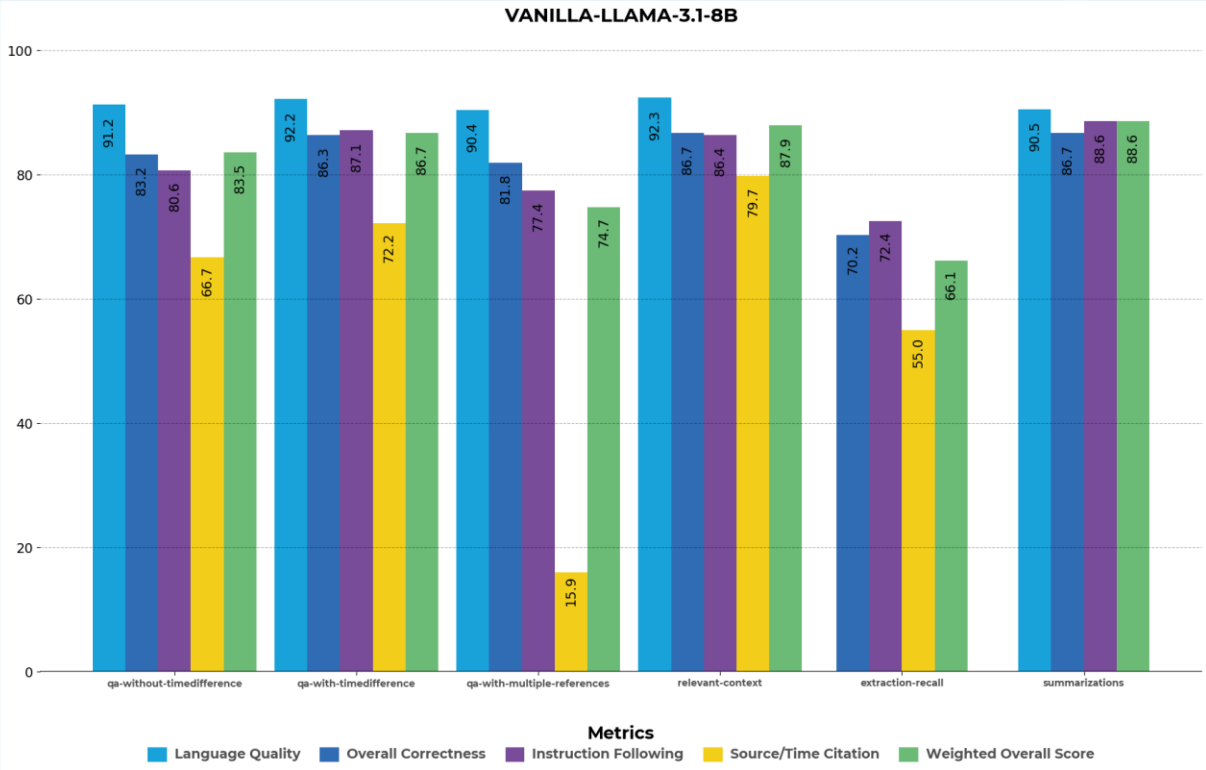
Compared to GPT-3.5-Turbo, this MERGED model still demonstrates superior performance in source citations and time awareness. The overall weighted scores remain better than vanilla in most categories (50-88%), except for the noted decreases in extraction recall.

This suggests MISTRAL-7B-MERGED is not matching the expectation set from the other MERGED-Model Results and even decreased in Performance after Merging compared to ORPO or SFT. We advise to NOT USE the MISTRAL-7B-MERGED Model and see it as an experiment. We will investigate on that more in the future.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**140,85**	**2,86**
qa-with-timedifference	**121,66**	**5,96**
qa-with-multiple-references	**546,96**	**1,62**
relevant-context	**-10,94**	**-12,19**
extraction-recall	**-49,30**	**-42,28**
summarizations	**/**	**-24,71**

*Differences compared to GPT-3.5-Turbo

EASY-BENCHMARK | Evaluation of LLAMA-3.1-8B



Comparing Llama-3.1-8B with GPT-3.5-Turbo, the models show similar performance in several areas but with some notable differences:

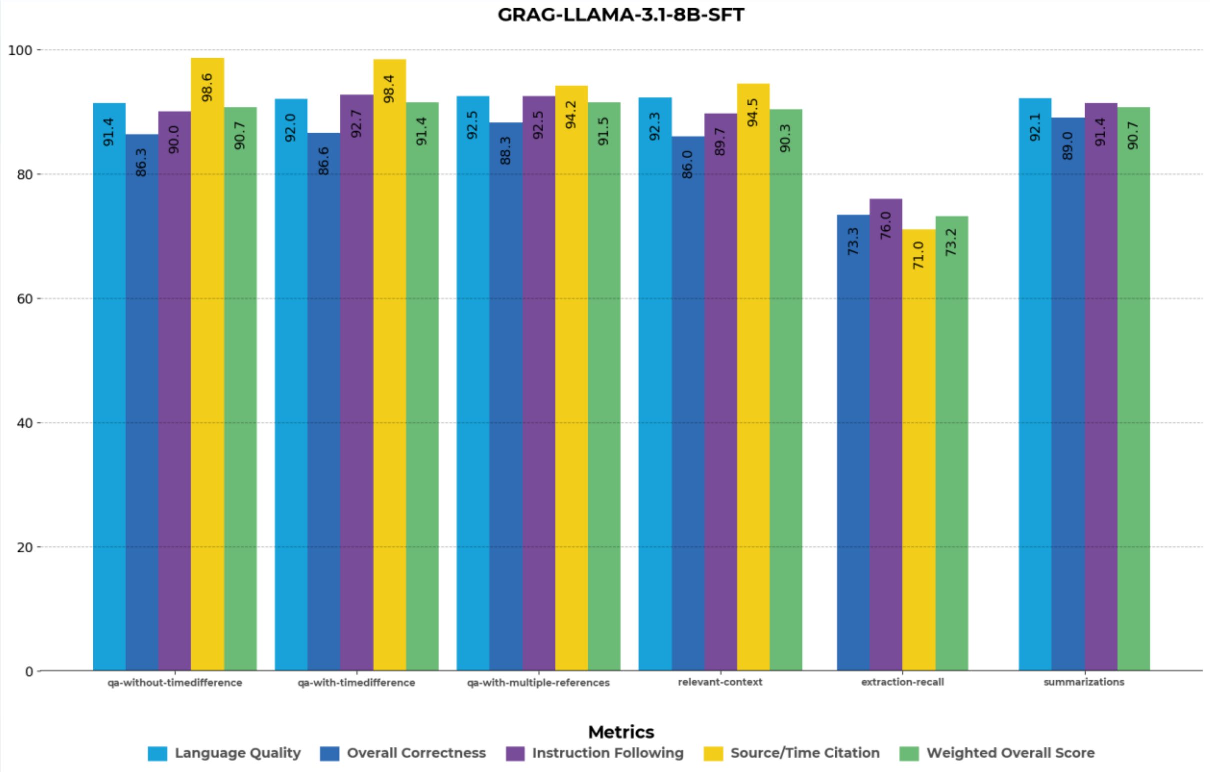
Language Quality and Overall Correctness are comparable, with both models maintaining high scores (85-95%) across most tasks. However, Llama shows slightly better source and time citations in basic tasks **(66-72% vs GPT-3.5's 39-43%)** but performs slightly worse overall with multiple references.

Notable differences appear in extraction-recall tasks, where Llama-3.1-8B scores lower **(around 55% vs GPT-3.5's 84.5%)**. Both models perform similarly well in summarization tasks, suggesting Llama-3.1-8B has strong fundamental capabilities in these areas despite being a smaller model.

The weighted overall scores remain comparable between both models, typically ranging from 75-85%, indicating that Llama-3.1-8B provides competitive performance against GPT-3.5-Turbo even before any fine-tuning.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**67,17**	**0,51**
qa-with-timedifference	**66,36**	**4,13**
qa-with-multiple-references	**72,83**	**-3,24**
relevant-context	**10,39**	**-1,80**
extraction-recall	**-34,96**	**-24,20**
summarizations	**/**	**1,97**

*Differences compared to GPT-3.5-Turbo



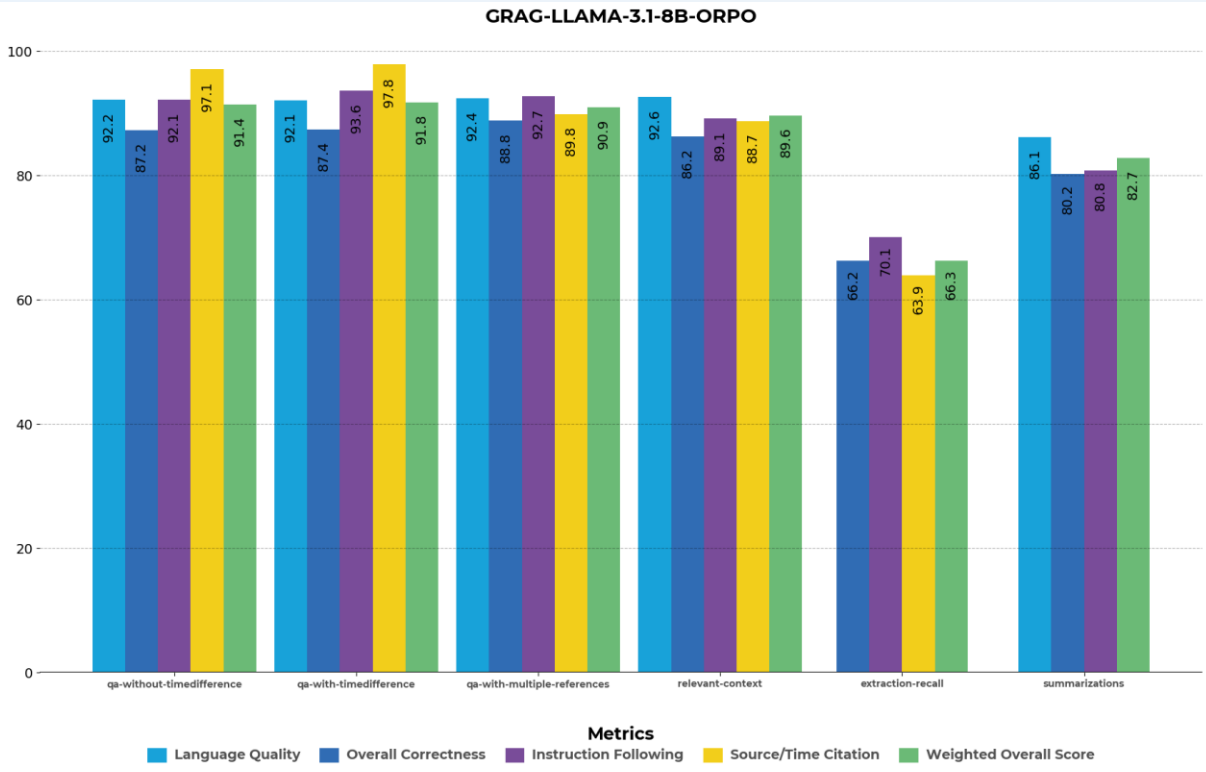
The GRAG-LLAMA-3.1-8B-SFT shows impressive improvements over both its vanilla version and GPT-3.5-Turbo. Most notably, the source and time citations have improved dramatically to over 96% accuracy, significantly outperforming GPT-3.5-Turbo's 39-43% range. The model also shows strong gains in source citing for multiple-references (94.2% vs GPT-3.5's 9.2%).

Language quality remains consistently high (~90%) across all tasks, matching GPT-3.5-Turbo. The extraction-recall performance has improved from the vanilla version's 55% to about 73%, though still not quite reaching GPT-3.5-Turbo's 84.5%. Summarization capabilities remain comparable to GPT-3.5-Turbo, with weighted overall scores consistently above 90% for most tasks.

The most significant achievement is the model's superior performance in source attribution and temporal awareness while maintaining strong performance across other metrics. This suggests that our SFT training has successfully specialized the model for German RAG tasks without compromising its general capabilities, often exceeding GPT-3.5-Turbo's performance despite being 14 times smaller model.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**147,12**	**9,17**
qa-with-timedifference	**126,73**	**9,88**
qa-with-multiple-references	**923,91**	**18,50**
relevant-context	**30,89**	**0,97**
extraction-recall	**-15,95**	**-16,07**
summarizations	**/**	**4,47**

*Differences compared to GPT-3.5-Turbo



Source and Time Citations remain excellent but show slight decreased variations:

- QA without time difference: 97.1% (ORPO) vs. 98.4% (SFT)
- QA with time difference: 97.8% (ORPO) vs. 98.6% (SFT)
- Multiple references: 89.8% (ORPO) vs. 94.2% (SFT)

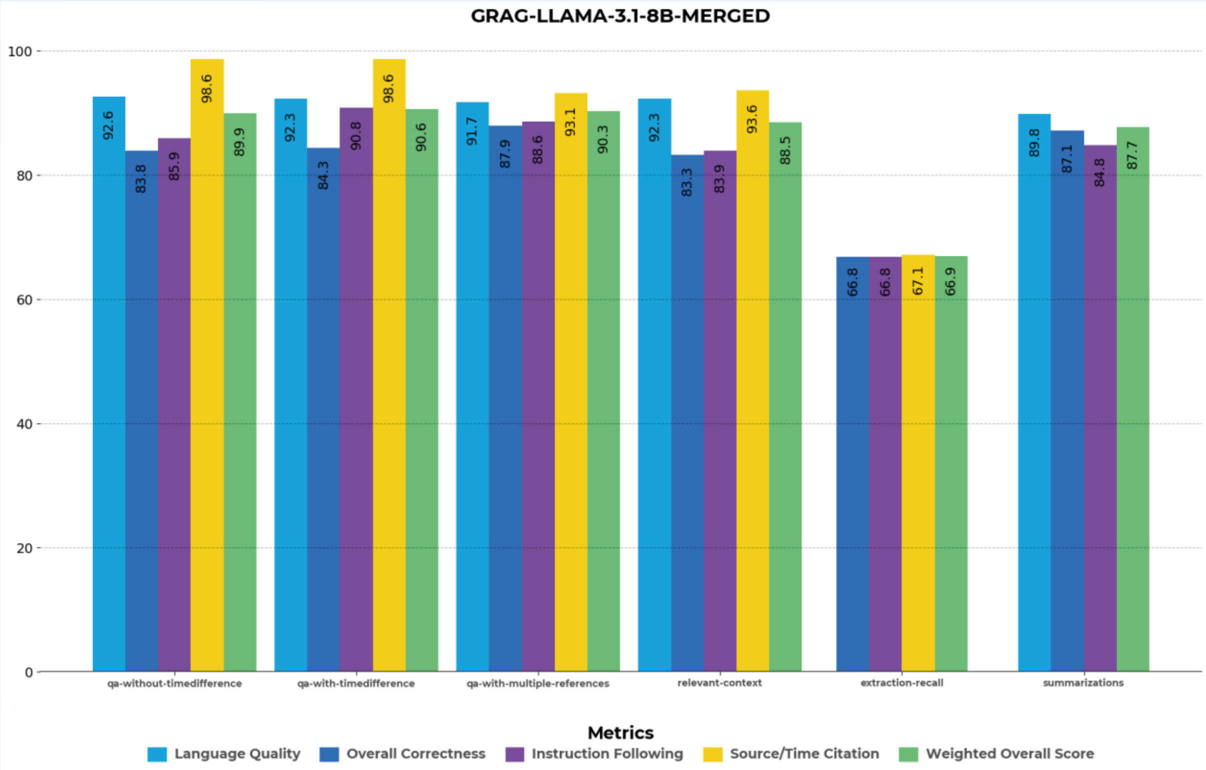
Notable Decreases:

- Extraction-recall drops from 73.3% to 66.2%
- Overall correctness shows slight decreases across several tasks
- Language quality remains similar but with minor fluctuations

Despite these changes, the MERGED model maintains a strong weighted overall score across most tasks, though slightly lower than the SFT version. This suggests that while the ORPO training may have introduced some trade-offs, it hasn't significantly degraded the model's core capabilities established during SFT training.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**143,33**	**9,99**
qa-with-timedifference	**125,35**	**10,28**
qa-with-multiple-references	**876,09**	**17,84**
relevant-context	**22,85**	**0,12**
extraction-recall	**-24,38**	**-23,97**
summarizations	**/**	**-4,78**

*Differences compared to GPT-3.5-Turbo



Source and Time Citations remain excellent and slightly increased compared to ORPO:

- QA without time difference: **98,6%** (MERGED) vs. **97.1%** (ORPO) vs. **98.4%** (SFT)
- QA with time difference: **98,6%** (MERGED) vs. **97.8%** (ORPO) vs. **98.6%** (SFT)
- Multiple references: **93,1%** (MERGED) vs. **89.8%** (ORPO) vs. **94.2%** (SFT)
- Relevant Context: **93,6%** (MERGED) vs. **88.7%** (ORPO) vs. **94.5%** (SFT)
- Extraction Recall: **67,1%** (MERGED) vs. **63.9%** (ORPO) vs. **71.0%** (SFT)

Notable Decreases:

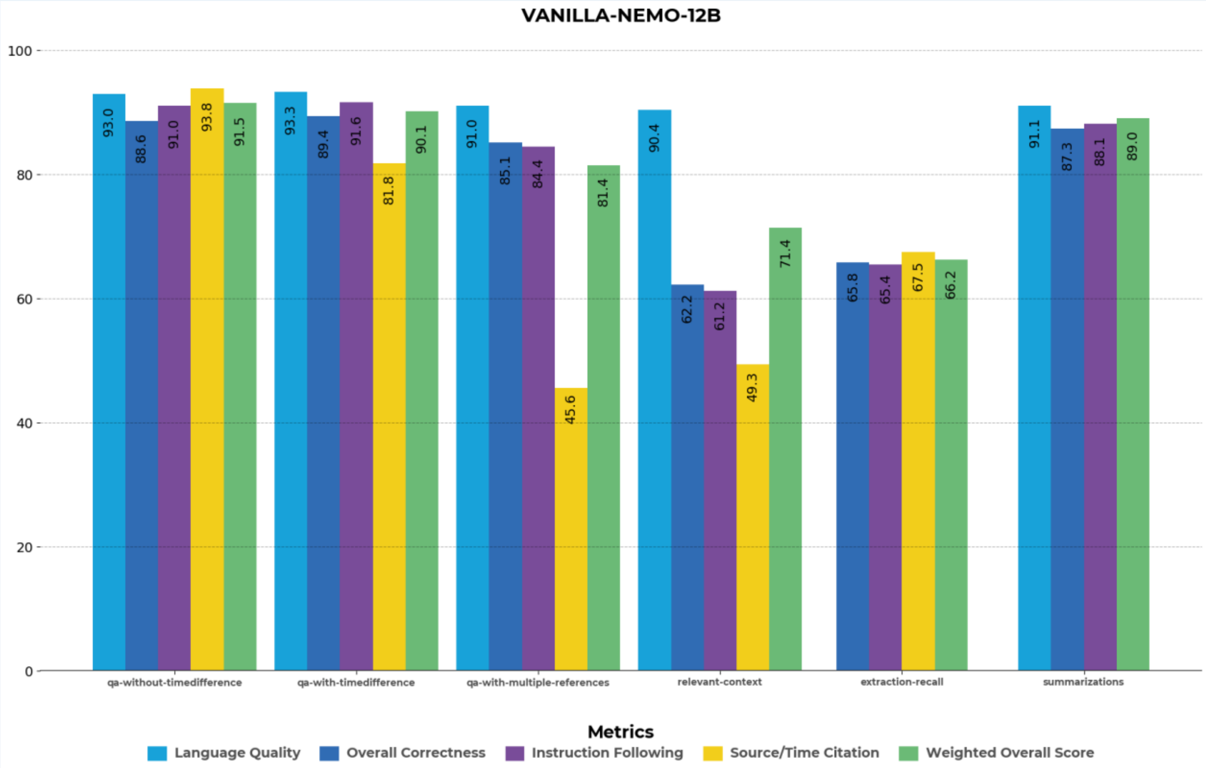
- Overall correctness shows slight decreases across several tasks
- Language quality remains similar but with minor fluctuations

Despite these changes, the ORPO model maintains a strong weighted overall score across most tasks, though slightly lower than the SFT version. This suggests that while the ORPO training may have introduced some trade-offs, it hasn't significantly degraded the model's core capabilities established during SFT training.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**147,12**	**8,12**
qa-with-timedifference	**127,19**	**8,92**
qa-with-multiple-references	**911,96**	**16,97**
relevant-context	**29,64**	**-1,08**
extraction-recall	**-20,59**	**-23,27**
summarizations	**/**	**0,97**

*Differences compared to GPT-3.5-Turbo

EASY-BENCHMARK | Evaluation of NEMO-12B



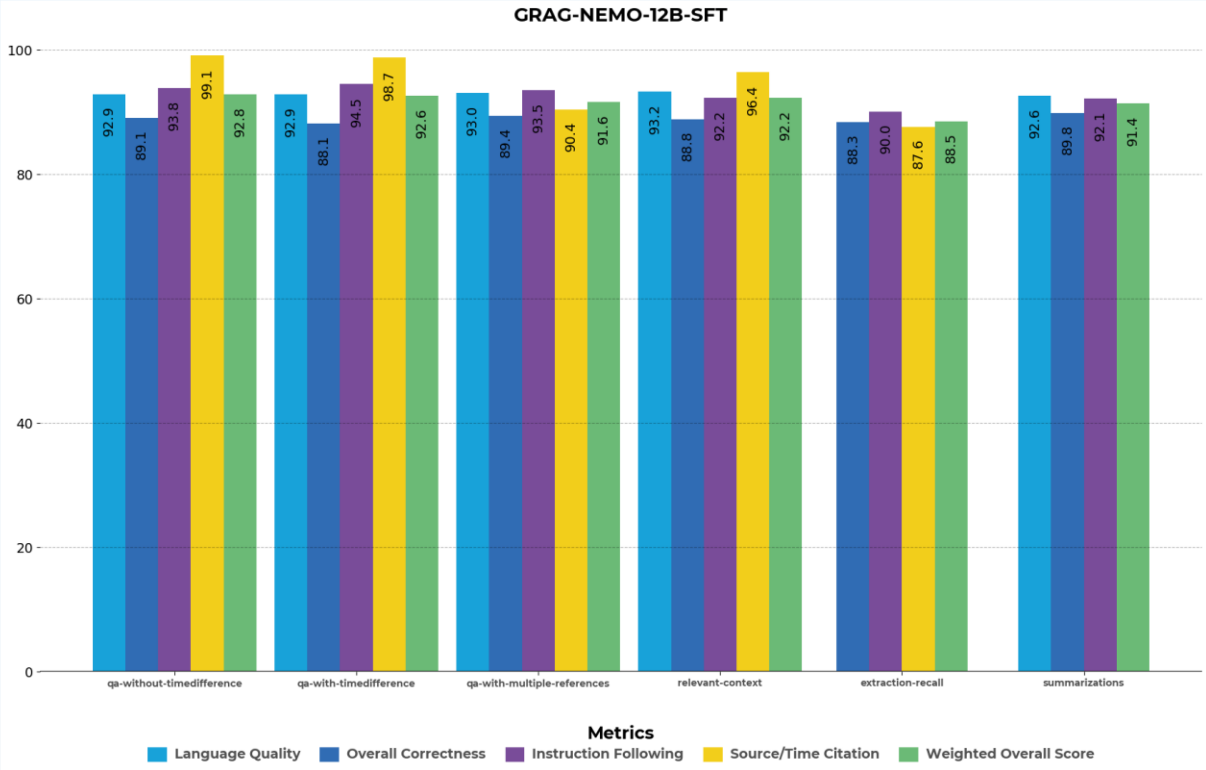
NEMO-12B shows some interesting performance patterns compared to GPT-3.5-Turbo, despite being a larger model compared to the before trained Models:

- QA without time difference: 93.9% vs GPT-3.5's 39.9%
- QA with time difference: 81.8% vs GPT-3.5's 43.4%
- However, it still struggles with multiple references (45.6% vs GPT-3.5's 9.82%)
- Language Quality is comparable between both models, with both achieving around 90-95% across tasks.
- Relevant context handling (62.2% vs GPT-3.5's 91.3%)
- Extraction-recall (65.8% vs GPT-3.5's 87.5%)

The weighted overall scores remain fairly similar, though NEMO-12B shows slightly lower performance in some areas.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**135,09**	**10,06**
qa-with-timedifference	**88,48**	**8,32**
qa-with-multiple-references	**395,65**	**5,52**
relevant-context	**-31,72**	**-20,21**
extraction-recall	**-20,12**	**-24,04**
summarizations	**/**	**2,44**

*Differences compared to GPT-3.5-Turbo



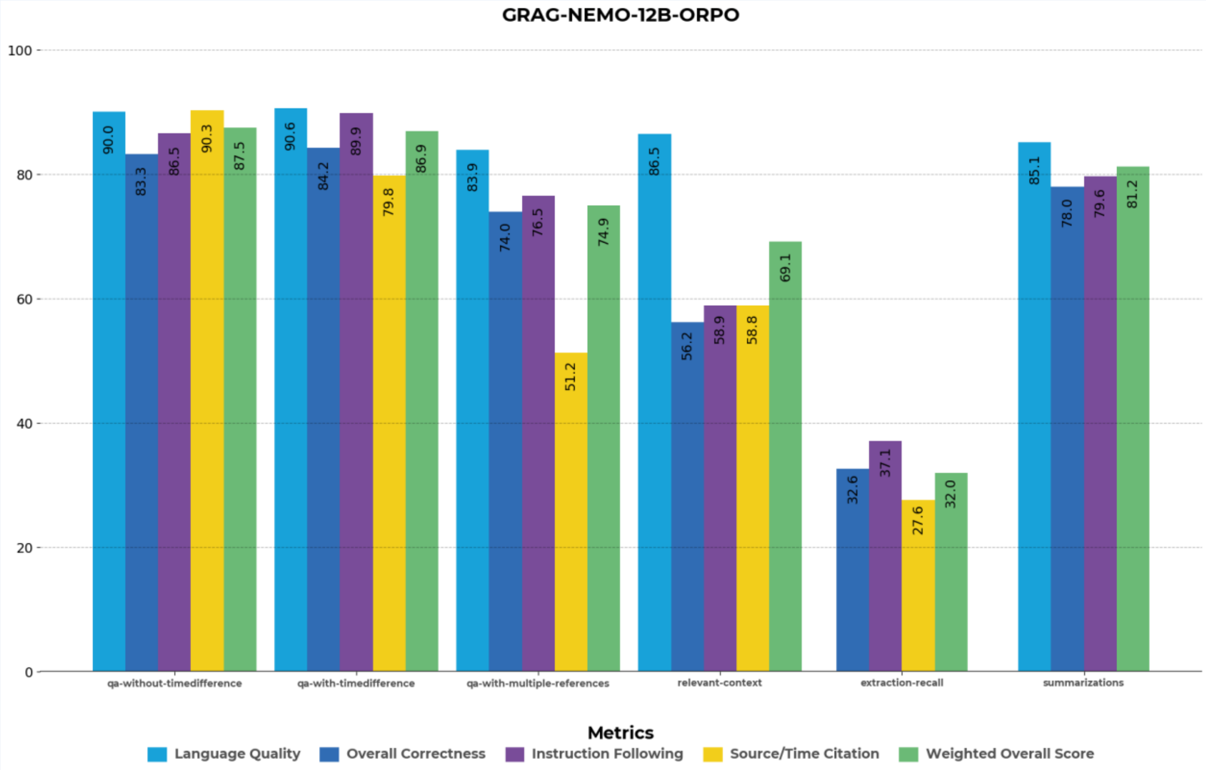
The GRAG-NEMO-12B-SFT demonstrates remarkable improvements over both its vanilla version and GPT-3.5-Turbo across all metrics. Most notably, the model's handling of source and time citations has been dramatically enhanced, maintaining accuracy rates above 98% compared to GPT-3.5-Turbo's 39-43% range. The most impressive improvement is seen in multiple reference handling, where the model jumped from the vanilla version's 45.6% to 94.4%, far surpassing GPT-3.5-Turbo's 9.82%.

The SFT training has also significantly improved areas where the vanilla version struggled. Relevant context handling improved from 71,4% to 92.5%, while extraction-recall capabilities increased from 66.2% to 88.5%. The model now exceeds GPT-3.5-Turbo's performance in these areas while maintaining strong language quality and instruction following capabilities.

Weighted overall scores consistently remain above 90% across most tasks, indicating that the SFT training has successfully leveraged NEMO-12B's larger capacity. This success demonstrates the effectiveness of our training approach in creating a specialized model for German RAG applications.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**148,37**	**11,65**
qa-with-timedifference	**127,42**	**11,22**
qa-with-multiple-references	**882,61**	**18,68**
relevant-context	**33,52**	**3,07**
extraction-recall	**3,68**	**1,46**
summarizations	**/**	**5,19**

*Differences compared to GPT-3.5-Turbo



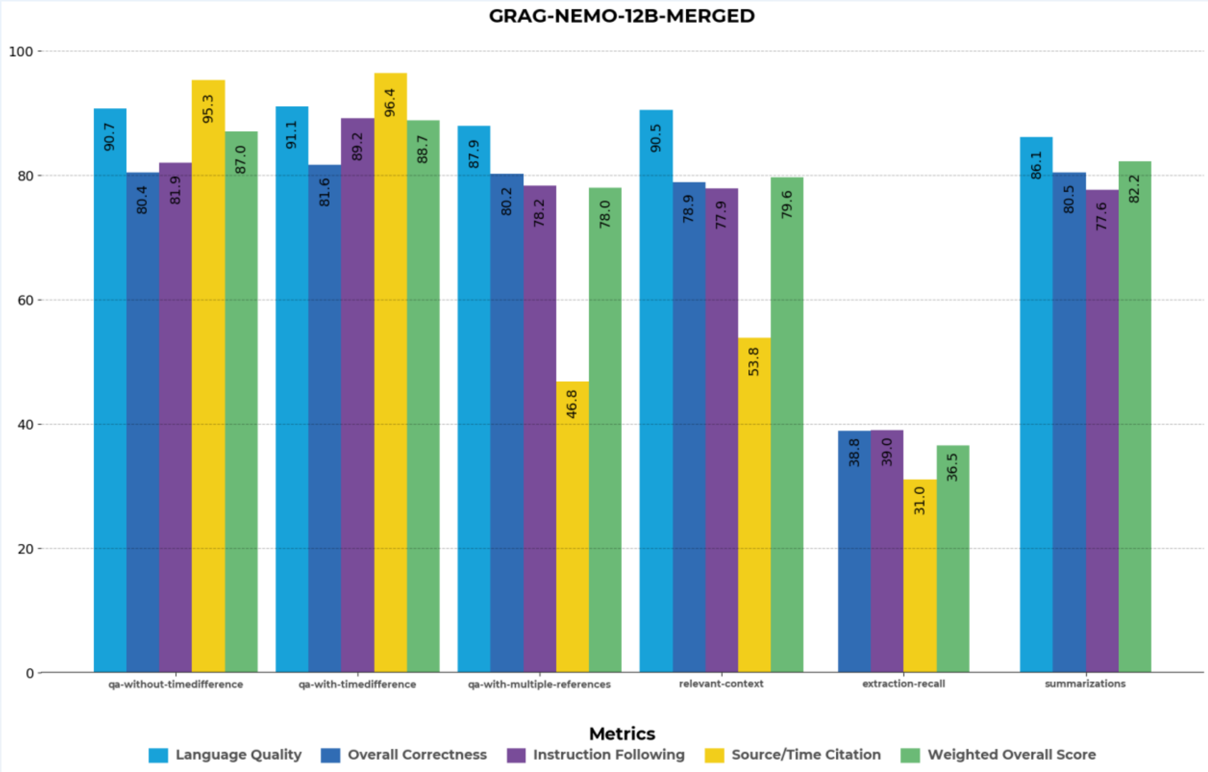
After applying ORPO training to the NEMO-12B model, we observe some significant changes compared to the SFT version. The model shows slight decreases in several key performance areas, most notably in the extraction-recall task where performance dropped dramatically from 88.5% to 32.0%. This represents a concerning regression in one of the model's core capabilities.

The source and time citation capabilities, while still strong, have decreased from their previous peaks of 98-99% to around 90% for single references and 79.8% for time-aware tasks. Multiple reference handling has also declined from 94.4% to 51.2%, though this still remains significantly better than GPT-3.5-Turbo's baseline performance.

While the language quality remains consistently high across most tasks, the overall weighted scores show a general decline across tasks. This suggests that the ORPO training phase, while intended to improve the model's alignment and instruction following, may have introduced some unintended trade-offs. Because of that even for this bigger Models merging the SFT- & ORPO-Model seemed to be a good approach to us.

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	**126,29**	**5,28**
qa-with-timedifference	**83,87**	**4,49**
qa-with-multiple-references	**456,63**	**-2,95**
relevant-context	**-18,53**	**-22,76**
extraction-recall	**-67,38**	**-63,33**
summarizations	**/**	**-6,55**

*Differences compared to GPT-3.5-Turbo



After **MERGING** the **NEMO-12B SFT- & ORPO** model, we observe some significant changes compared to the Checkpoints before. The model shows slight decreases in several key performance areas, most notably in the extraction-recall task where performance dropped dramatically from 88.3% to 32.6%. This represents a concerning regression in one of the model's core capabilities.

The model sees a noticeable drop in multiple-references (down to around 46% from 51%), relevant context (down to around 53% from 58%) suggesting that the MERGING may have introduced alarming trade-offs in these areas. However, it still maintains acceptable performance in core RAG tasks like question answering with / without time differences.

Compared to GPT-3.5-Turbo, this MERGED model still demonstrates superior performance in source citations and time awareness. The overall weighted scores decreased compared to vanilla in a few task categories (36-88%).

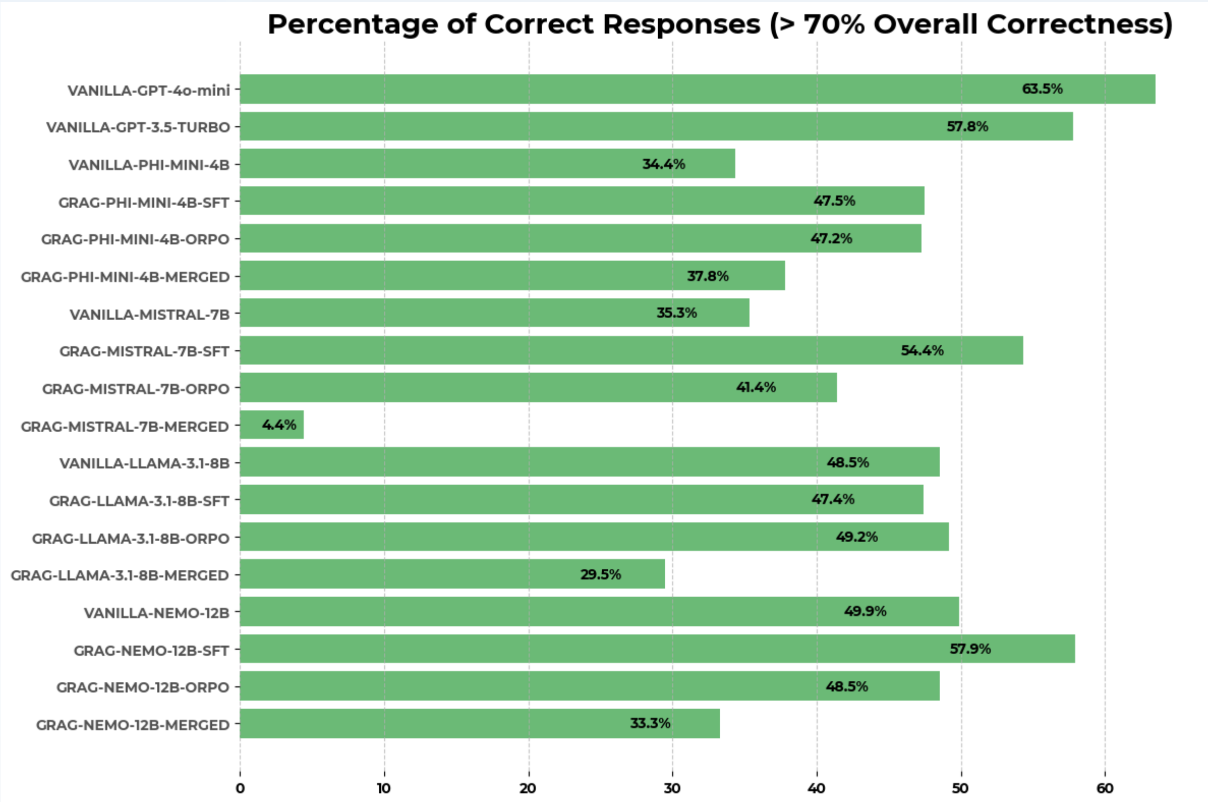
This suggests NEMO-12B-MERGED, again a Model from Mistral, is not matching the expectation set from the other MERGED-Model Results and even decreased in Performance after Merging compared to ORPO or SFT Checkpoints. **We advise to NOT USE the NEMO-12B-MERGED Model and see it as an experiment. We will investigate on that more in the future.**

Task	Difference Citation (%)*	Difference Overall (%)*
qa-without-timedifference	138,85	4,64
qa-with-timedifference	122,12	6,65
qa-with-multiple-references	408,70	1,01
relevant-context	-25,48	-11,02
extraction-recall	-63,31	-58,12
summarizations	/	-5,41

*Differences compared to GPT-3.5-Turbo

EASY-BENCHMARK | Evaluation: Reasoning

This section compares the performance of our models with their corresponding instruct base models. The first plot illustrates the percentage of correct responses across the models. As mentioned, GPT-4o is utilized as the evaluator to determine whether the responses generated by the models are correct. As shown in the figure, Mistral models demonstrate notable performance. In particular, **GRAG-NEMO-SFT** achieves a slightly higher score than **GPT-3.5-Turbo**, despite being a 12-billion-parameter model and significantly smaller than GPT-3.5. Moreover, Mistral-7B-SFT demonstrates significant improvement compared to Mistral-7B-Instruct, LLAMA-8B-SFT, and LLAMA-8B-Instruct, indicating that the Mistral models are better adapted to the reasoning subset. Overall, we believe the results are promising, especially considering that the performance gap between **GRAG-NEMO** and **GPT-4o-mini** is relatively small, despite GRAG-NEMO being significantly smaller in size.



As reported before the Performance Degradation of the **MERGED Mistral Models** after Merging leads to our advice to **NOT USE** these Models and see them as experiments. We also have seen strange repetation of non-sense tokens when using the “standard” system instructions from summarization & reasoning tasks “Folge den Anweisungen des Benutzers …” (translated: “Follow the Instructions of the User…”), that could only be “resolved” by prepending an additional sentence in the system instruction. But also this change has not accomplished a real fix for these models so far.

EASY-BENCHMARK | Evaluation: OCR & JSON

This section compares the performance of our models with their corresponding instruct base models as also GPT-3.5-turbo & GPT-4o-mini for the tasks of correcting OCR received texts and generating valid JSON that adheres to key names and ENUM values for each key.

The **"Generate valid JSON"** table represents the overall score for the checks of valid key-names, valid key-values and valid JSON-Syntax. Only when all the of the checks are passed each evaluated row get an overall score of 100. This is then divided by the number of successfully evaluated rows to get the final overall score per model. The Goal of the Training Dataset was not to adhere the models to a predefined classification tasks that could infer the model outputs in the future. The goal was to prime the models to generate valid JSON syntax and follow constrains when selecting the values for each key. The trained examples could also be possible to classify as other categories and types so in this evaluation it is not important to check for the "target-values" as strict indicators.

The **"Correcting OCR-received Texts"** table represents the Word Error Rates per Model and the Difference between the Model WER & the Input WER (how many Word Errors were inputted into the input text). The errors were generated synthetically by a script on synthetically generated wikipedia article summarizations following the findings of **"Scrambled text: training Language Models to correct OCR errors using synthetic data".**

PHI-Models

Generate valid JSON

Model Name	Overall Score
GRAG-PHI-MINI-4B-MERGED	34.8
GRAG-PHI-MINI-4B-ORPO	**79.8**
GRAG-PHI-MINI-4B-SFT	68.4
VANILLA-PHI-MINI-4B	72.5
VANILLA-GPT-3.5-TURBO	**99.9**
VANILLA-GPT-4o-mini	**100**

Correcting OCR-received Texts

Model Name	Model WER	Difference Input WER
GRAG-PHI-MINI-4B-MERGED	**0.12**	**-0.09**
GRAG-PHI-MINI-4B-ORPO	1.27	1.06
GRAG-PHI-MINI-4B-SFT	0.31	0.1
VANILLA-PHI-MINI-4B	1.01	0.8
VANILLA-GPT-3.5-TURBO	**0.06**	**-0.15**
VANILLA-GPT-4o-mini	**0.05**	**-0.16**

MISTRAL-Models

Generate valid JSON

Model Name	Overall Score
GRAG-MISTRAL-7B-MERGED	**51.8**
GRAG-MISTRAL-7B-ORPO	8
GRAG-MISTRAL-7B-SFT	47.1
VANILLA-MISTRAL-7B	12.5
VANILLA-GPT-3.5-TURBO	**99.9**
VANILLA-GPT-4o-mini	**100**

Correcting OCR-received Texts

Model Name	Model WER	Difference Input WER
GRAG-MISTRAL-7B-MERGED	0.98	0.77
GRAG-MISTRAL-7B-ORPO	1.13	0.92
GRAG-MISTRAL-7B-SFT	**0.05**	**-0.16**
VANILLA-MISTRAL-7B	4.82	4.61
VANILLA-GPT-3.5-TURBO	**0.06**	**-0.15**
VANILLA-GPT-4o-mini	**0.05**	**-0.16**

LLAMA-Models

Generate valid JSON

Model Name	Overall Score
GRAG-LLAMA-3.1-8B-MERGED	87.3
GRAG-LLAMA-3.1-8B-ORPO	0
GRAG-LLAMA-3.1-8B-SFT	**99.2**
VANILLA-LLAMA-3.1-8B	13.5
VANILLA-GPT-3.5-TURBO	**99.9**
VANILLA-GPT-4o-mini	**100**

Correcting OCR-received Texts

Model Name	Model WER	Difference Input WER
GRAG-LLAMA-3.1-8B-MERGED	4.65	4.44
GRAG-LLAMA-3.1-8B-ORPO	3.36	3.15
GRAG-LLAMA-3.1-8B-SFT	**0.1**	**-0.1**
VANILLA-LLAMA-3.1-8B	1.16	0.95
VANILLA-GPT-3.5-TURBO	**0.06**	**-0.15**
VANILLA-GPT-4o-mini	**0.05**	**-0.16**

NEMO-Models

Generate valid JSON

Model Name	Overall Score
GRAG-NEMO-12B-MERGED	**99.5**
GRAG-NEMO-12B-ORPO	79.8
GRAG-NEMO-12B-SFT	**99.6**
VANILLA-NEMO-12B	76.9
VANILLA-GPT-3.5-TURBO	**99.9**
VANILLA-GPT-4o-mini	**100**

Correcting OCR-received Texts

Model Name	Model WER	Difference Input WER
GRAG-NEMO-12B-MERGED	2.39	2.18
GRAG-NEMO-12B-ORPO	1.27	1.06
GRAG-NEMO-12B-SFT	**0.05**	**-0.15**
VANILLA-NEMO-12B	5.54	5.33
VANILLA-GPT-3.5-TURBO	**0.06**	**-0.15**
VANILLA-GPT-4o-mini	**0.05**	**-0.16**

Summarization of the Evaluation of OCR & JSON

Interestingly the trends we have seen in the evaluated SFT-Tasks before, when applying Preference Training with ORPO or the Performance Degradation after Merging can not be fully confirmed for these tasks. One of the most surprising changes were that the merged Mistral Models (Mistral-7B & Nemo-12B) again performed quite well on the JSON-Task. Also, the merged Phi-Model sees the first real performance decrease in SFT-Tasks for generating valid JSON but achieves way better results than the other Checkpoints in OCR-Correction. The ORPO-Model from Llama totally forgot the task to generate valid JSON which was recovered again by merging the SFT- & ORPO-Model but hasn't got back to previous performance of the SFT-Model **(87% vs. 99%)**. Also, the Merging of Llama has not recovered its ability of correcting OCR-received texts to an acceptable Rate. In Comparison to GPT-3.5-turbo and GPT-4o-mini all the GRAG-Models get to a comparable reliability in Terms of generating valid JSON and correcting OCR-received Texts, except the Phi-Models and the Mistral-7B Model in terms of generating valid JSON.

EASY-BENCHMARK | Evaluation: Overview

Retrieval Augmented Tasks

Comparison of the weighted Overall-Score across best performing Models per Base-Model-Type for working and generating with context-based Answers:

Task	PHI-MERGED	MISTRAL-SFT	LLAMA-SFT	NEMO-SFT	GPT-3.5-TURBO	GPT-4o-mini
qa-without-timedifference	88,01	91,38	90,74	92,80	83,12	92,46
qa-with-timedifference	89,12	92,01	91,43	92,55	83,21	93,23
qa-with-multiple-references	84,77	91,17	91,46	91,60	77,18	88,72
extraction-recall	61,76	91,68	73,17	88,45	87,18	67,34
relevant-context	84,40	89,65	90,34	92,22	89,47	90,66
summarizations	84,89	89,53	90,74	91,37	86,86	91,30

Generate valid JSON

Model Name	Overall Score
GRAG-PHI-MINI-4B-ORPO	79.8
GRAG-MISTRAL-7B-MERGED	51.8
GRAG-LLAMA-3.1-8B-SFT	99.2
GRAG-NEMO-12B-SFT	99.6
VANILLA-GPT-3.5-TURBO	99.9
VANILLA-GPT-4o-mini	100

Correcting OCR-received Texts

Model Name	Model WER	Difference Input WER
GRAG-PHI-MINI-4B-MERGED	0.12	-0.09
GRAG-MISTRAL-7B-SFT	0.05	-0.16
GRAG-LLAMA-3.1-8B-SFT	0.1	-0.1
GRAG-NEMO-12B-SFT	0.05	-0.15
VANILLA-GPT-3.5-TURBO	0.06	-0.15
VANILLA-GPT-4o-mini	0.05	-0.16

Based on these results you can choose the best Base-Model for your RAG-specific Task by adhering to the provided Prompt-Templates or Fine-Tune these Models on your domain-specific Context or comparable Task-Descriptions further.

IMPORTANT: We would not advice to use the GRAG-MISTRAL-7B-MERGED or GRAG-NEMO-12B-MERGED in any production Use-Case because they seem to be instable.

HARD-BENCHMARK | Evaluations

Approach & Weighting

This section presents the performance of our models on the HARD-BENCHMARK dataset, comparing them with GPT models and the Instruct-tuned versions of our models. Additionally, it introduces a comprehensive evaluation framework that combines rigorous automated metrics with advanced LLM-based assessments to provide a thorough analysis of model performance.

For the qualitative aspects of evaluation, we employ an LLM-as-judge approach, where a separate language model (**gpt-4o-mini**) evaluates responses based on specific criteria. This judging model is prompted with carefully designed rubrics to assess language quality, instruction adherence, and factual accuracy. The LLM judge processes each response independently, providing numerical scores and detailed justifications for each metric. This approach allows for consistent evaluation of subjective aspects while maintaining reproducibility across assessments.

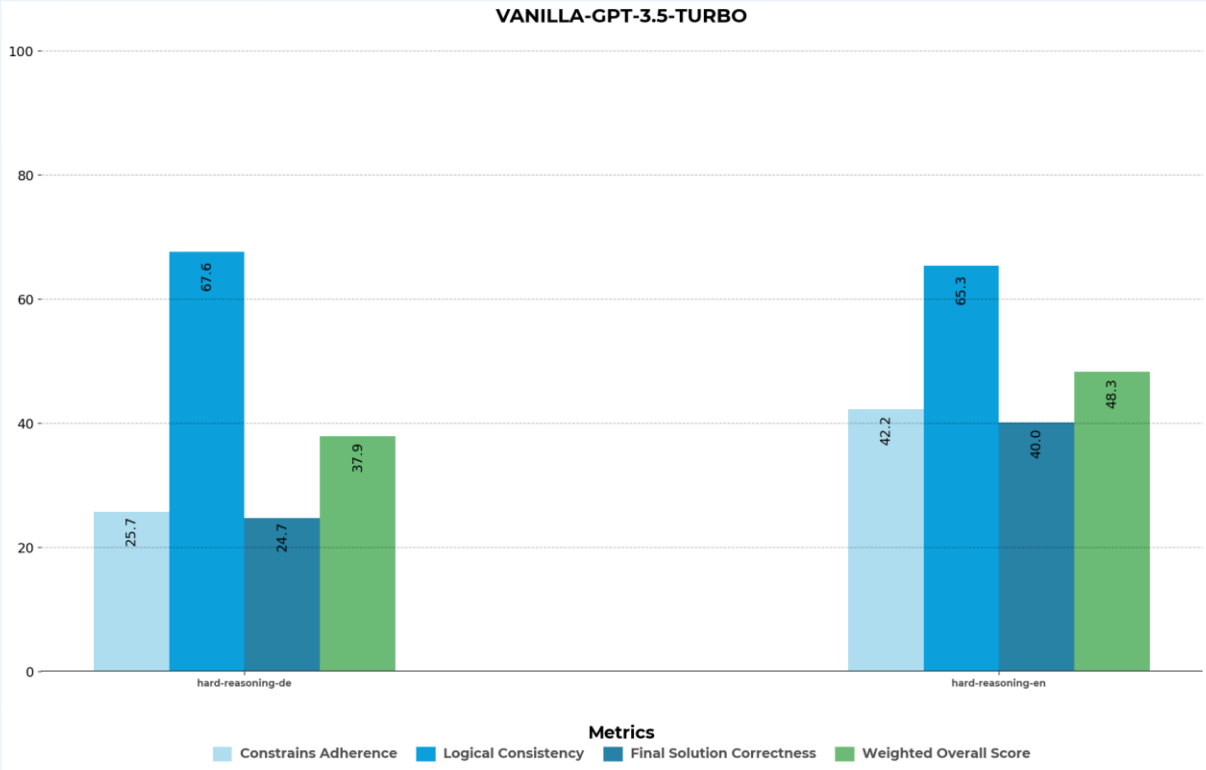
The final weighted score combines qualitative measurements, with predefined weight given to LLM-based assessments. This balanced approach ensures that models are evaluated not only on their technical accuracy but also on the qualitative aspects of their outputs that are crucial for real-world applications. This Evaluation **includes only 2 tasks that are trained with the same structure of instruction (hard-reasoning-de/en).** Also, we release 3 not yet trained eval-subsets for harder multiple-references examples and meeting specific summarizations which asks for detailed summaries of specific topics mentioned from all attendees or from specific attendees, but we were not able to get reliable outputs for evals from GRAG or Vanilla-Models which caused the exclusion of these subsets for this evaluation part of the report.

In conclusion, each GRAG model is compared with its corresponding instruct model, and GPT models with plots representing a detailed analysis of the following metrics:

- **Constraints Adherence**
- **Logical Consistency**
- **Final Solution Correctness**
- **Weighted Overall score**

To calculate the **Weighted Overall Score** metric, we employed a method that prioritizes the **Final Solution Correctness** metric. If this metric exceeds **70**, the **Weighted Overall Score** is calculated using a specific weighting strategy. However, for cases where the Final Solution Correctness metric is below **70**, a **Weighted Overall Score** of zero is assigned. The employed weighting strategy is represented as follows:

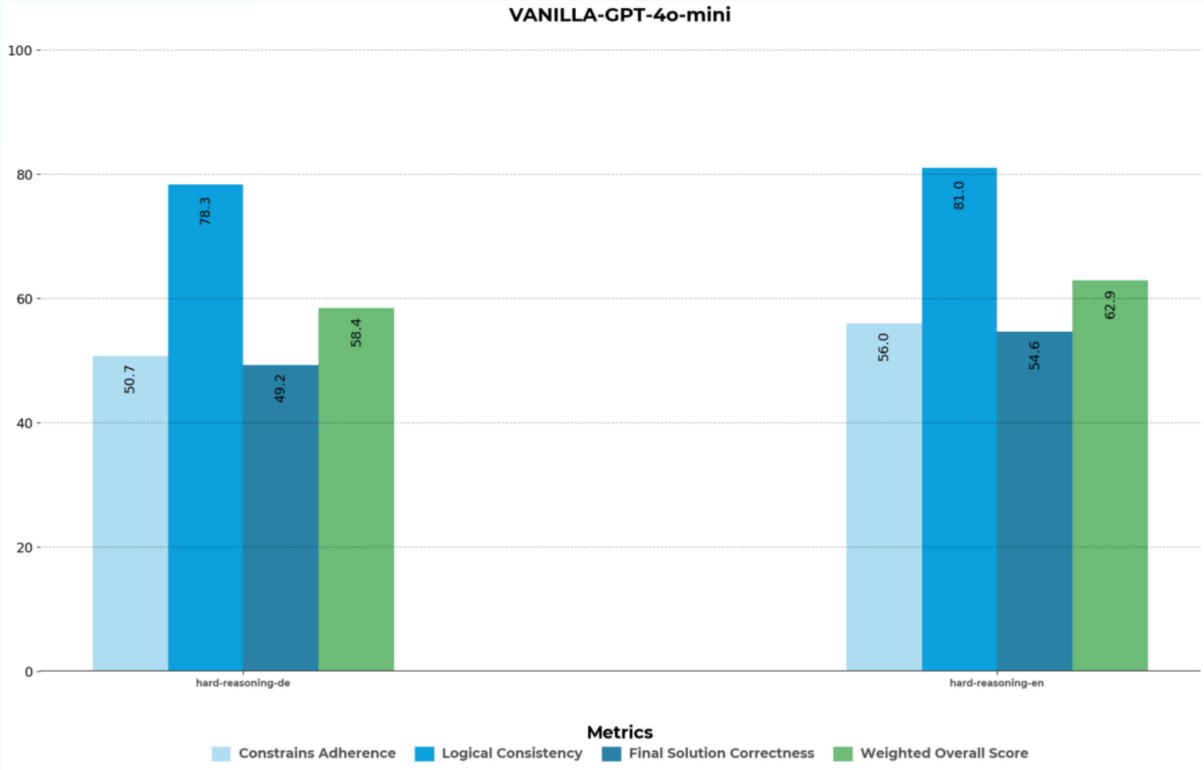
Metric	Weights
Constrains Adherence	**0.3**
Logical Consistency	**0.3**
Final Solution Correctness	**0.4**



GPT-3.5-Turbo's performance on hard reasoning tasks reveals notable disparities between German and English processing. In German reasoning, while maintaining strong logical consistency (67.6%), the model shows weak constraint adherence (25.7%) and final solution correctness (24.7%), resulting in a modest weighted score of 37.9%.

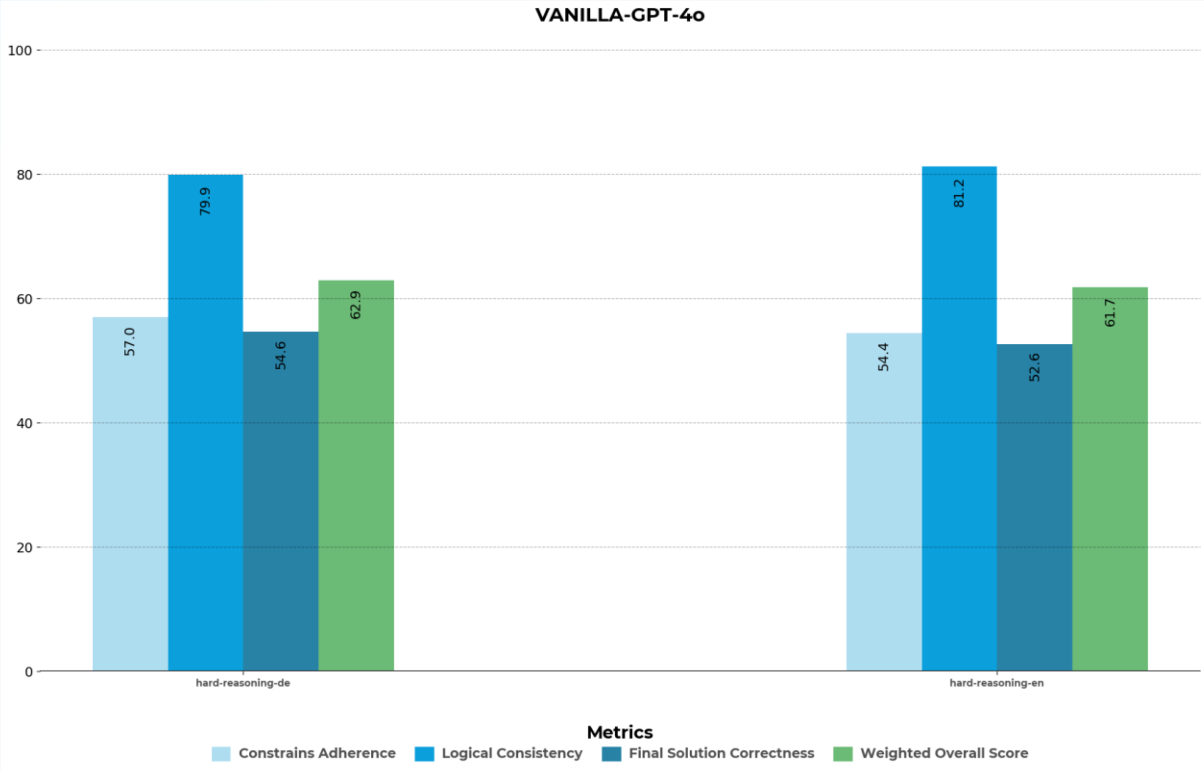
English reasoning shows marginally better results with improved constraint adherence (42.2%), logical consistency (65.3%), and final solution correctness (40.0%), achieving a weighted score of 48.3%. However, both scores fall significantly below our 70% validity threshold.

This performance gap between languages, despite GPT-3.5-Turbo's extensive training, highlights persistent challenges in multilingual reasoning capabilities, particularly in constraint management for non-English tasks.



GPT-4o-mini demonstrates substantial improvements over GPT-3.5-Turbo in hard reasoning tasks across both languages. In German reasoning, the model shows strong performance in logical consistency at 78.3%, with constraint adherence reaching 50.7%. The final solution correctness of 49.2% contributes to an overall strong weighted score of 58.4%.

The model performs even better on English reasoning tasks, where it achieves higher constraint adherence at 56.0% and exceptional logical consistency at 81.0%. With a final solution correctness of 54.6%, it achieves an impressive weighted score of 62.9%. The consistent improvement across all metrics, particularly in logical consistency, showcases GPT-4o-mini's enhanced reasoning capabilities compared to GPT-3.5-Turbo, while also demonstrating better balance between constraint adherence and solution correctness.

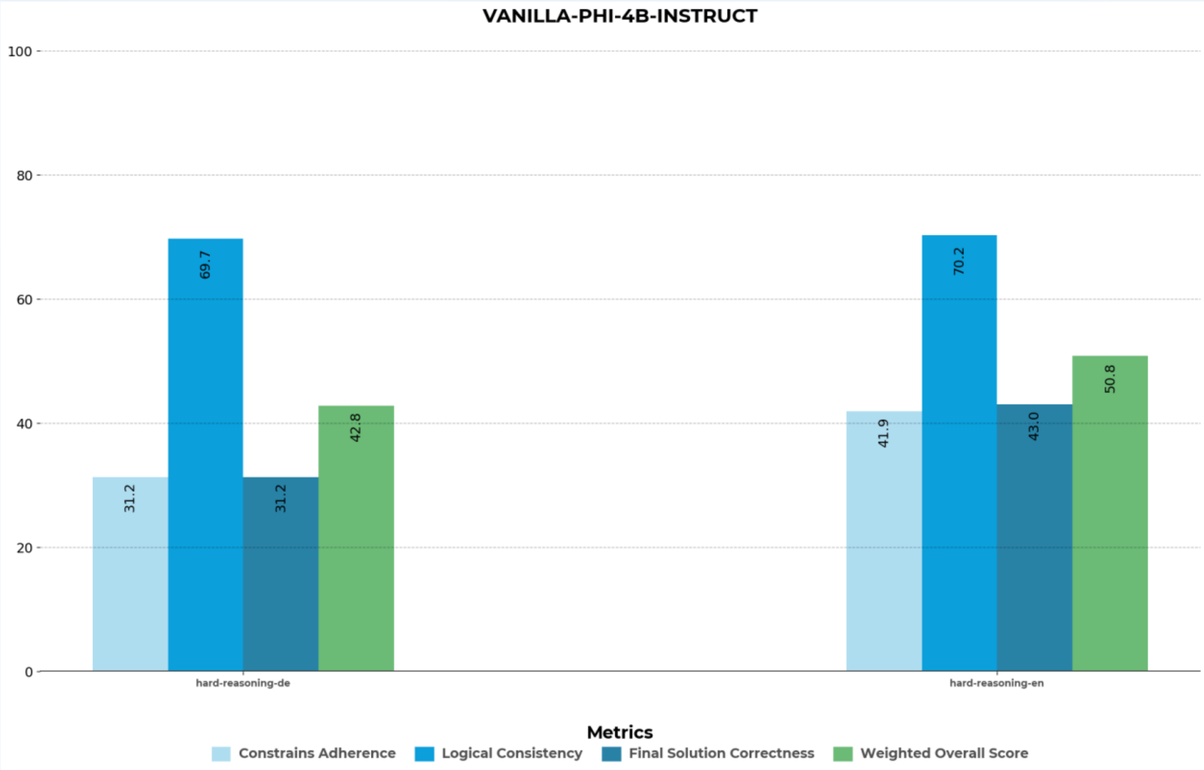


GPT-4o shows remarkable performance in hard reasoning tasks, building on the strengths seen in GPT-4o-mini while achieving even better consistency. In German reasoning, the model demonstrates impressive constraint adherence at 57.0% and maintains excellent logical consistency at 79.9%. With a final solution correctness of 54.6%, it achieves a strong weighted score of 62.9%.

The model's performance in English reasoning follows a similar pattern of excellence. It maintains strong constraint adherence at 54.4% and achieves outstanding logical consistency at 81.2%. Though the final solution correctness slightly decreases to 52.6%, the model still achieves a robust weighted score of 61.7%.

What's particularly interesting is how GPT-4o maintains more balanced scores between German and English reasoning compared to earlier models, suggesting improved multilingual reasoning capabilities. The high logical consistency scores in both languages, coupled with strong constraint adherence, indicate that the model has developed sophisticated reasoning capabilities that translate well across languages. The slight performance difference between German and English has narrowed significantly compared to GPT-3.5-Turbo, demonstrating more equitable handling of multilingual reasoning tasks.

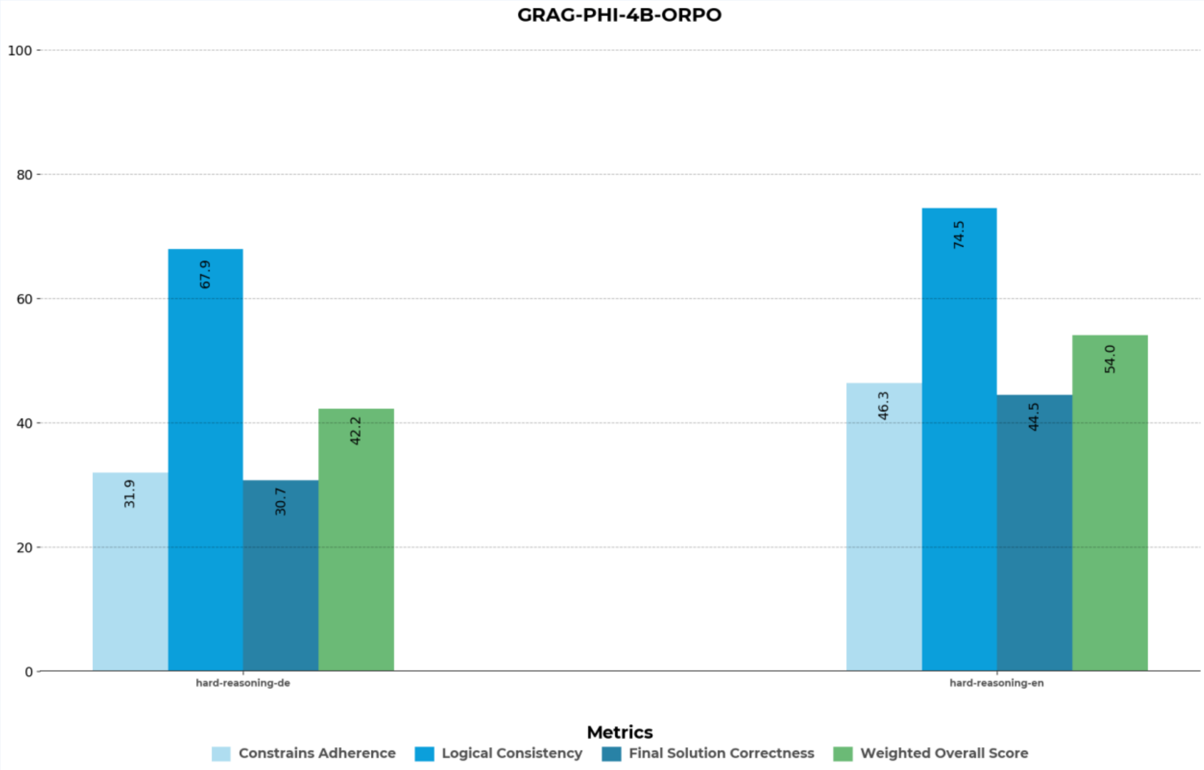
HARD-BENCHMARK | Evaluation of PHI-3.5-MINI



Looking at Microsoft's PHI-4B-INSTRUCT performance on hard reasoning tasks reveals interesting capabilities for a model of its size (4 billion parameters). In German reasoning tasks, while constraint adherence is limited at 31.2%, the model demonstrates surprisingly strong logical consistency at 69.8%. With a final solution correctness matching its constraint adherence at 31.2%, it achieves a weighted score of 42.8%.

The model performs notably better in English reasoning tasks, showing improved constraint adherence at 41.9% and maintaining strong logical consistency at 70.2%. The final solution correctness rises to 43.0%, contributing to a weighted score of 50.8%.

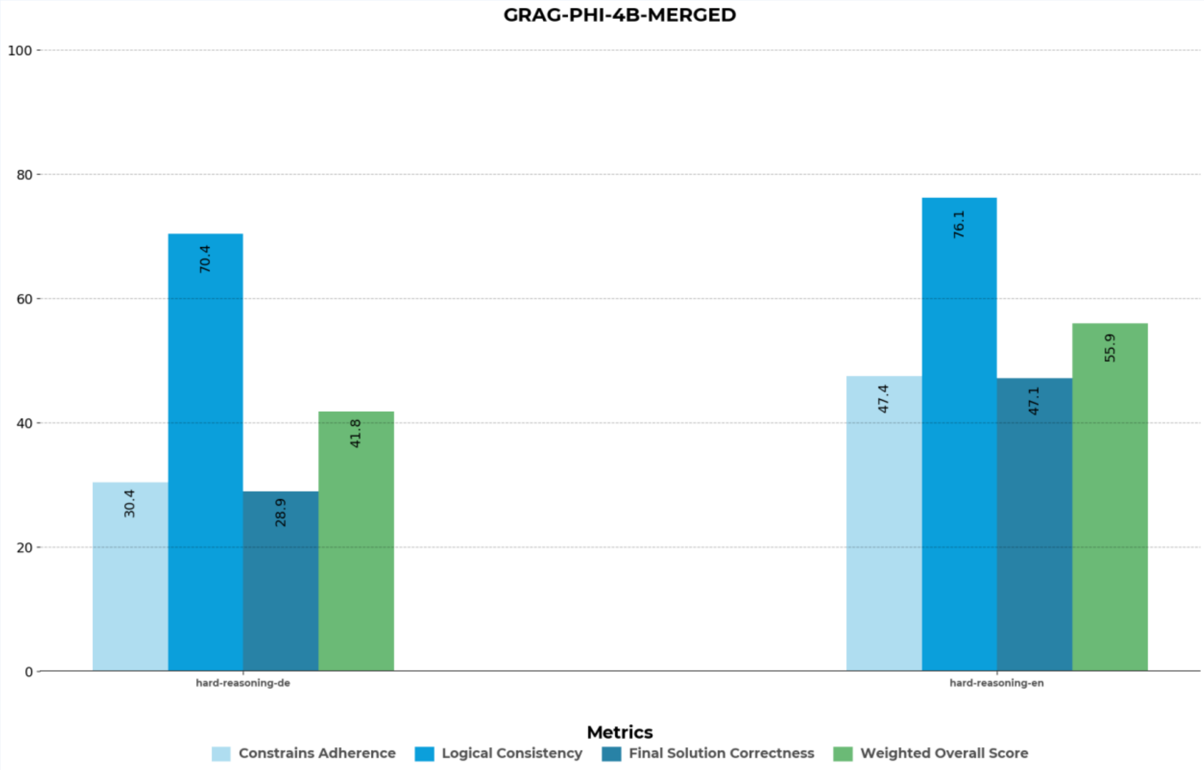
What makes these results particularly interesting is that despite being significantly smaller than GPT-3.5-Turbo (175B parameters), PHI-4B-INSTRUCT achieves comparable and sometimes better logical consistency scores. While its constraint adherence and solution correctness scores are lower than larger models like GPT-4o, its ability to maintain strong logical reasoning capabilities with just 4B parameters demonstrates the potential of efficient model architectures when properly instruction-tuned. This suggests that raw model size isn't the only determining factor in developing strong reasoning capabilities.



Our GRAG-PHI-4B-ORPO model demonstrates encouraging results after applying our specialized training methodology. In German reasoning tasks, while maintaining similar constraint adherence at 31.9%, the model shows a noteworthy improvement in logical consistency, reaching 67.9%. Though the final solution correctness of 30.7% leads to a weighted score of 42.2%, what's particularly interesting is how the model preserves its reasoning capabilities through the ORPO training phase.

The English reasoning results show even more promising developments. The model achieves significantly improved constraint adherence at 46.3% and demonstrates impressive logical consistency at 74.5%. With a final solution correctness of 44.5%, it reaches a weighted score of 54.0%. This substantial improvement in English reasoning suggests that our training approach effectively enhances the model's capabilities while maintaining its computational efficiency.

What makes these results particularly noteworthy is that our 4B parameter model achieves these scores after training from a base model, comparing favorably with much larger commercial models like GPT-3.5-Turbo in some metrics, especially logical consistency. This suggests that our targeted training approach can help smaller models develop sophisticated reasoning capabilities, particularly in handling English language tasks, while maintaining reasonable performance in German reasoning despite the significant size difference compared to larger models.



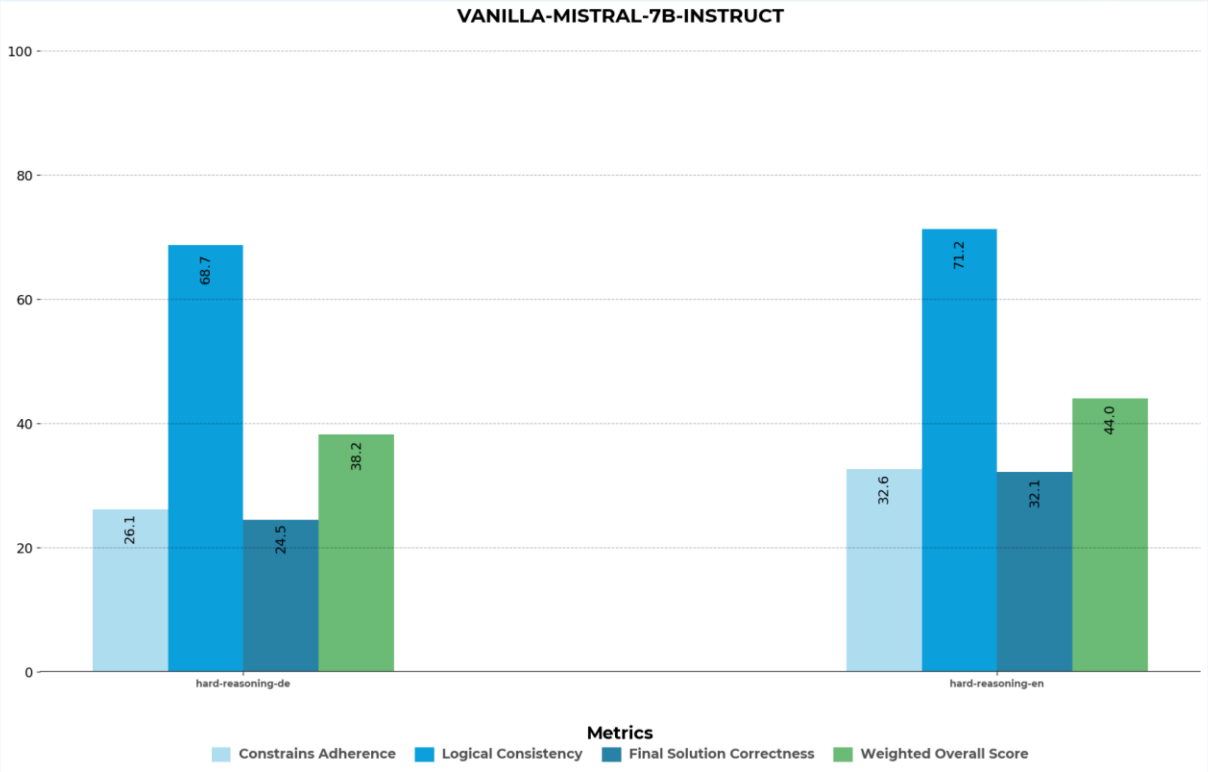
The GRAG-PHI-4B-MERGED model, which combines our SFT and ORPO training approaches, shows interesting developments in reasoning capabilities. In German reasoning tasks, while the constraint adherence remains modest at 30.4%, the model maintains strong logical consistency at 70.4%. The final solution correctness of 28.9% contributes to a weighted score of 41.8%, showing a slight trade-off in performance compared to individual training approaches.

For English reasoning tasks, the model demonstrates more robust performance. The constraint adherence improves significantly to 47.4%, while logical consistency reaches an impressive 76.1%. With a final solution correctness of 47.1%, the model achieves a weighted score of 55.9%. This enhanced English performance suggests that the merged training approach particularly benefits English language reasoning capabilities.

What makes these results especially interesting is how the merged approach affects different aspects of reasoning. While some metrics show slight decreases compared to the ORPO version, others improve, particularly in English tasks. This suggests that combining training approaches creates a more balanced model, though with some trade-offs between languages. The strong logical consistency scores, especially in English reasoning, indicate that the fundamental reasoning capabilities remain robust through the merging process, even as the model adjusts its handling of constraints and solution generation.

This performance pattern raises intriguing questions about how different training approaches interact and combine, particularly in the context of a smaller model architecture. The results suggest that while merging training approaches can enhance certain capabilities, careful consideration must be given to maintaining performance across all metrics and languages.

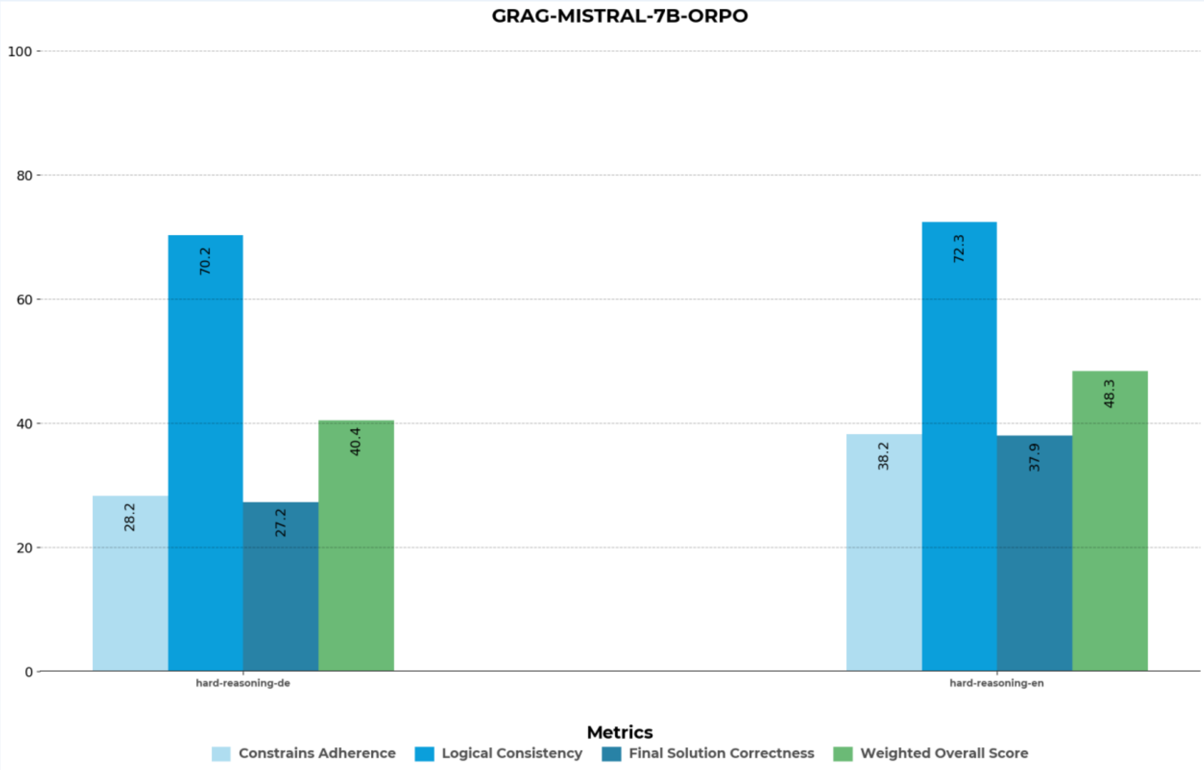
HARD-BENCHMARK | Evaluation of MISTRAL-7B



Mistral-7B-INSTRUCT demonstrates interesting capabilities as a mid-sized model in hard reasoning tasks. In German reasoning, the model shows moderate constraint adherence at 26.1% and achieves notable logical consistency at 68.7%. The final solution correctness of 24.5% leads to a weighted score of 38.2%. These scores indicate that even without specialized training, the model has developed meaningful reasoning capabilities in German language tasks.

The model performs somewhat better in English reasoning tasks, where it achieves higher constraint adherence at 32.6% and demonstrates strong logical consistency at 71.2%. With a final solution correctness of 32.1%, it reaches a weighted score of 44.0%. This performance gap between languages is common among models but remains relatively modest compared to some larger models.

What makes these results particularly interesting is that this is Mistral's instruction-tuned version, providing us with a valuable baseline for comparing our training approach. The model's strong logical consistency scores in both languages, despite being a 7B parameter model, suggest that Mistral's architecture efficiently handles complex reasoning tasks. This baseline performance helps us understand how our targeted training might further enhance these capabilities, particularly in areas like constraint adherence and solution correctness where there's clear room for improvement.



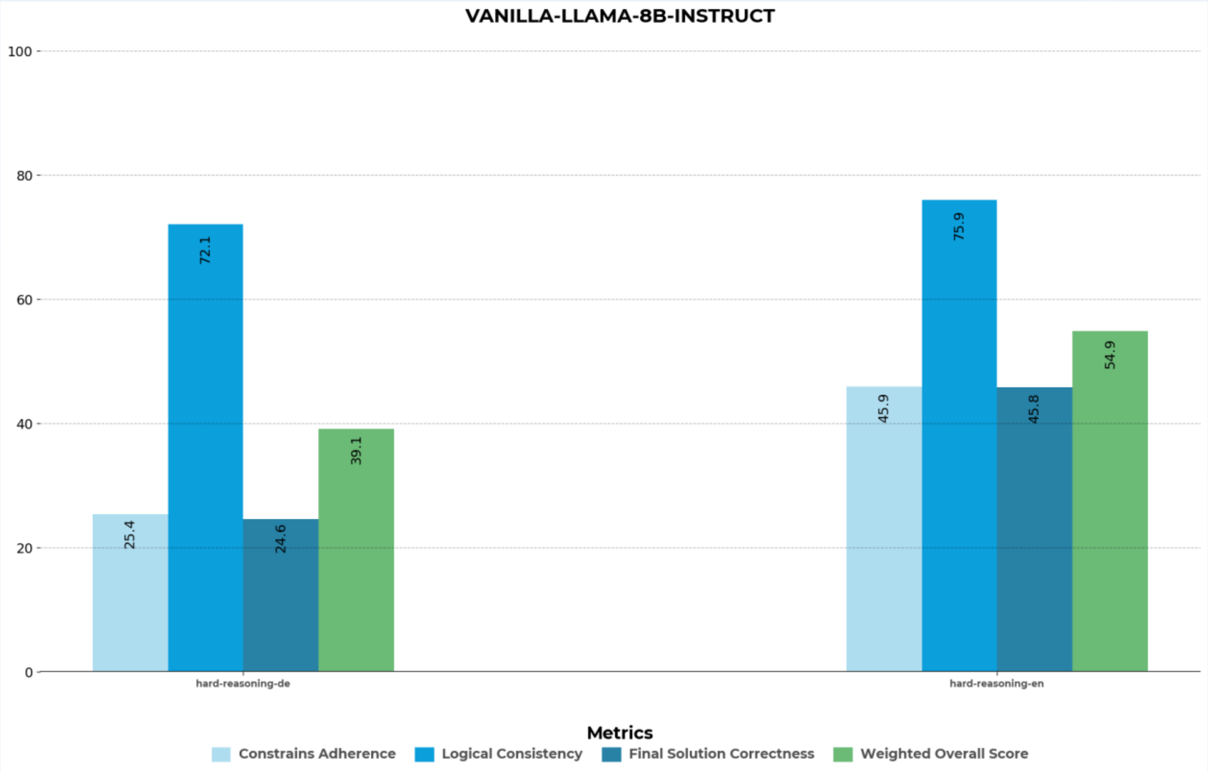
The GRAG-MISTRAL-7B-ORPO model presents an interesting alternative training approach to Mistral's instruction tuning. In German reasoning tasks, our model achieves constraint adherence of 28.2% and maintains strong logical consistency at 70.2%. With a final solution correctness of 27.2%, it reaches a weighted score of 40.4%. These results are comparable to Mistral's instruction-tuned version, with slightly different strengths in specific areas.

The English reasoning tasks show similar patterns of complementary capabilities. Our model demonstrates improved constraint adherence at 38.2% and maintains excellent logical consistency at 72.3%. The final solution correctness of 37.9% contributes to a weighted score of 48.3%. This performance suggests that different training approaches can achieve similar overall results through different paths.

What makes this comparison particularly fascinating is how different training methodologies can lead to comparable outcomes. While Mistral's instruction tuning achieves its results through broad-spectrum training, our ORPO approach reaches similar performance levels through targeted reasoning-focused training. This suggests that there might be multiple viable paths to developing strong reasoning capabilities in language models. The fact that both approaches achieve strong logical consistency scores while showing different patterns in constraint adherence and solution correctness provides valuable insights into how different training strategies affect various aspects of model performance.

These findings could be particularly valuable for the research community, as they demonstrate that effective model training isn't limited to a single approach. Both instruction tuning and our specialized training methodology can produce models with robust reasoning capabilities, each with its own unique strengths and characteristics.

HARD-BENCHMARK | Evaluation of LLAMA-3.1-8B

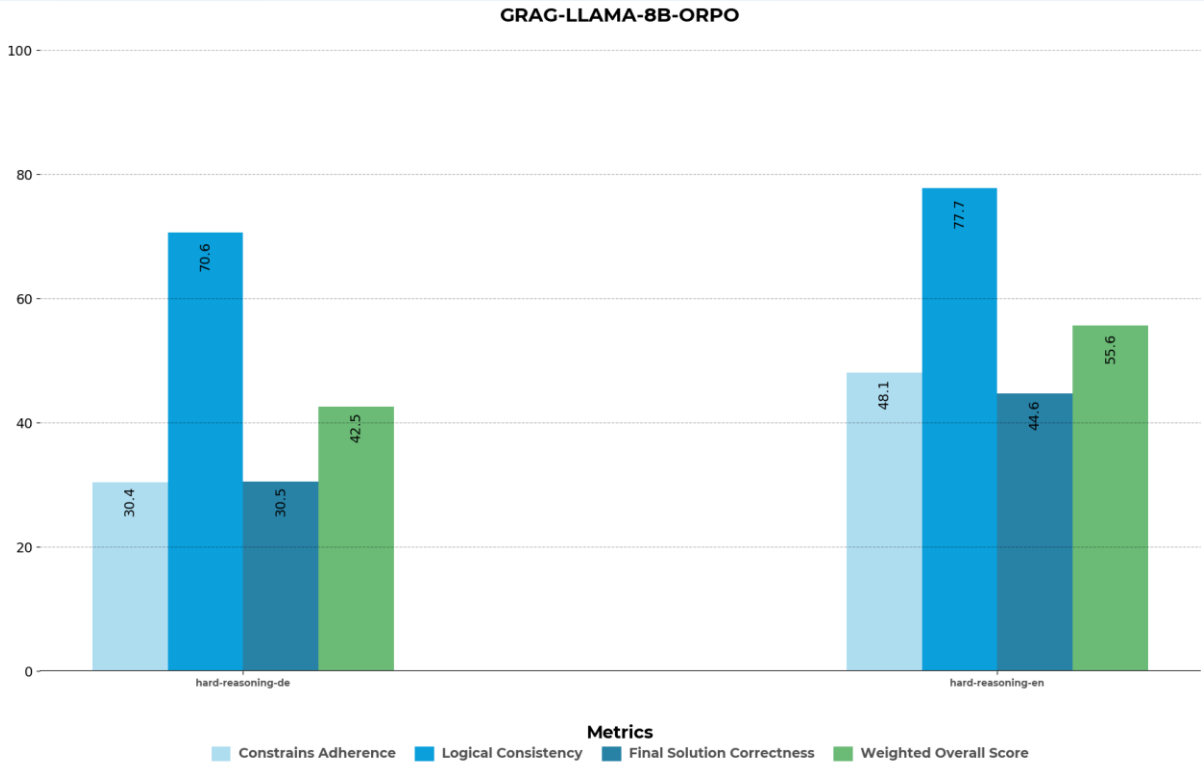


The Llama-3.1-8B-INSTRUCT model demonstrates Meta's instruction tuning capabilities in handling complex reasoning tasks. In German reasoning, the model shows moderate constraint adherence at 25.4% while achieving strong logical consistency at 72.1%. With a final solution correctness of 24.6%, it reaches a weighted score of 39.1%. These results highlight how instruction tuning can develop meaningful reasoning capabilities even in non-English languages.

In English reasoning tasks, the model shows more robust performance across all metrics. The constraint adherence improves significantly to 45.9%, while logical consistency reaches an impressive 75.1%. The final solution correctness of 45.8% contributes to a weighted score of 54.9%. This substantial improvement in English performance suggests that Meta's instruction tuning approach has been particularly effective for English language reasoning.

What makes these results especially interesting is that Meta's instruction tuning appears to create a strong foundation for logical consistency in both languages, even though other metrics show more variation between languages. The notably higher performance in English reasoning tasks provides valuable insights into how instruction tuning might be differently effective across languages. This performance pattern helps us understand both the strengths and potential areas for improvement in instruction-tuning approaches for multilingual models.

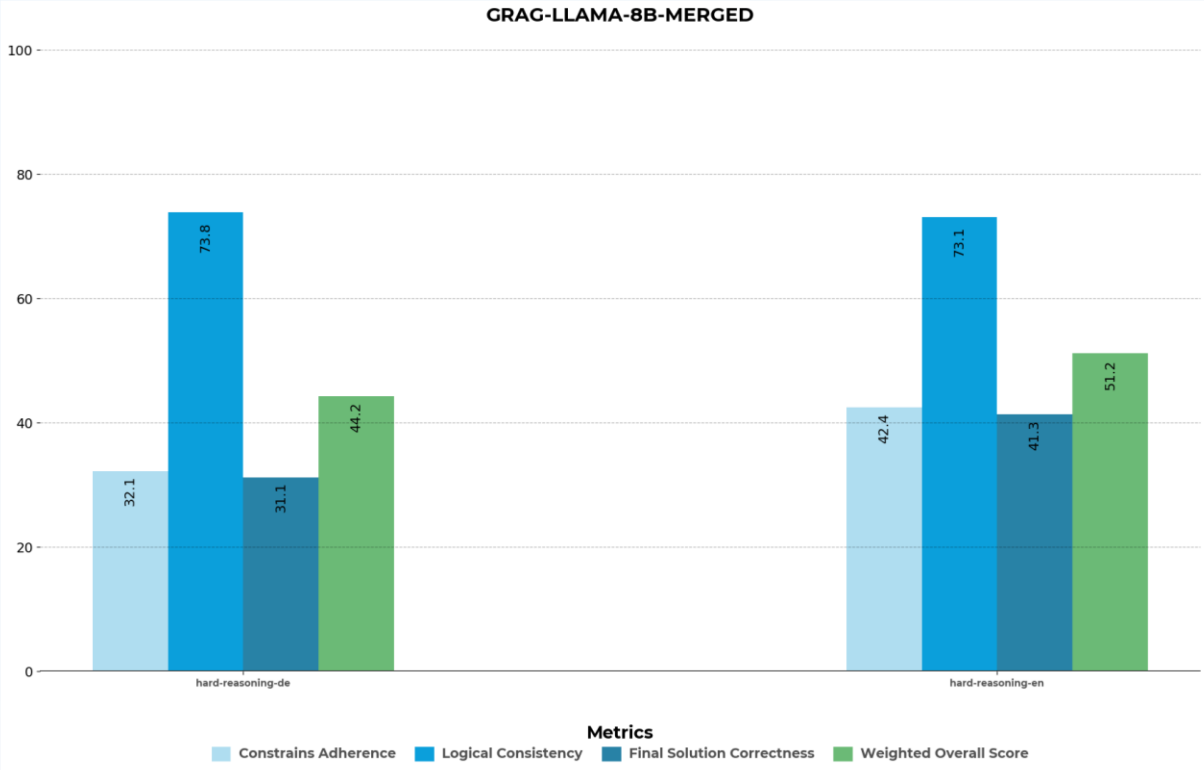
The considerable gap between German and English performance metrics also raises interesting questions about how instruction tuning might be optimized for better multilingual balance, particularly in areas like constraint adherence and solution correctness where we see the largest disparities between languages.



The GRAG-LLAMA-8B-ORPO model presents an alternative approach to developing reasoning capabilities compared to Meta's instruction tuning. In German reasoning tasks, our training methodology achieves constraint adherence of 30.4% and maintains strong logical consistency at 70.6%. With a final solution correctness of 30.5%, it reaches a weighted score of 42.5%. This showcases how different training philosophies can achieve comparable outcomes through distinct paths.

For English reasoning tasks, both approaches demonstrate strong capabilities with slightly different emphases. Our model shows constraint adherence of 48.1% and achieves impressive logical consistency at 77.7%. The final solution correctness of 44.6% leads to a weighted score of 55.6%. This suggests that both training methodologies can effectively develop sophisticated reasoning capabilities, though they may prioritize different aspects of performance.

What makes this comparison particularly enlightening is how it reveals different paths to developing reasoning capabilities in language models. While Meta's instruction tuning takes a comprehensive approach to developing general capabilities, our ORPO methodology focuses specifically on enhancing reasoning and output quality. The similar overall performance levels, achieved through different training strategies, suggest that multiple viable approaches exist for developing strong reasoning capabilities in language models. This insight could be valuable for the research community, as it demonstrates that effective model development isn't limited to a single training philosophy, but can be achieved through various carefully designed approaches, each with its own unique strengths and characteristics.

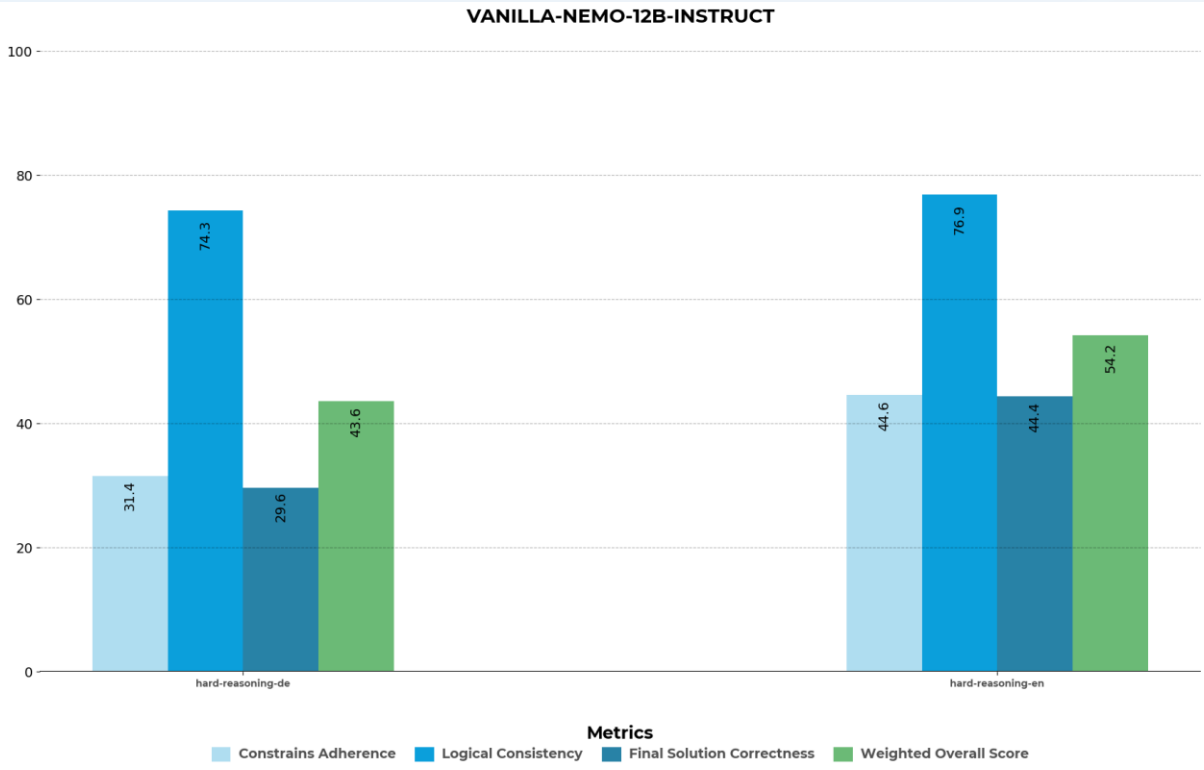


The GRAG-LLAMA-8B-MERGED model offers another interesting perspective on developing reasoning capabilities in language models. This version, which combines our different training approaches, shows distinct characteristics when compared to Meta's instruction-tuned version. In German reasoning tasks, the model demonstrates constraint adherence of 32.1% and maintains strong logical consistency at 73.8%. With a final solution correctness of 31.1%, it achieves a weighted score of 44.2%, showing how combining different training methodologies can yield unique performance patterns.

The English reasoning results tell an equally interesting story about different approaches to model development. Our merged version achieves constraint adherence of 42.4% and maintains excellent logical consistency at 73.1%. The final solution correctness of 41.3% contributes to a weighted score of 51.2%. When compared to Meta's instruction-tuned version, we see how different training philosophies can prioritize various aspects of reasoning capabilities while achieving similar overall performance levels.

This comparison provides valuable insights into the nature of language model training. Our merged approach combines targeted reasoning enhancement with output optimization. The fact that both approaches achieve comparable results through different paths suggests that the development of reasoning capabilities in language models isn't limited to a single methodology. Instead, we see how different training strategies can each contribute unique strengths to model performance, potentially opening up new avenues for research in model development and training optimization. This diversity in effective training approaches could be particularly valuable for researchers exploring different ways to enhance model capabilities while working within various computational and resource constraints.

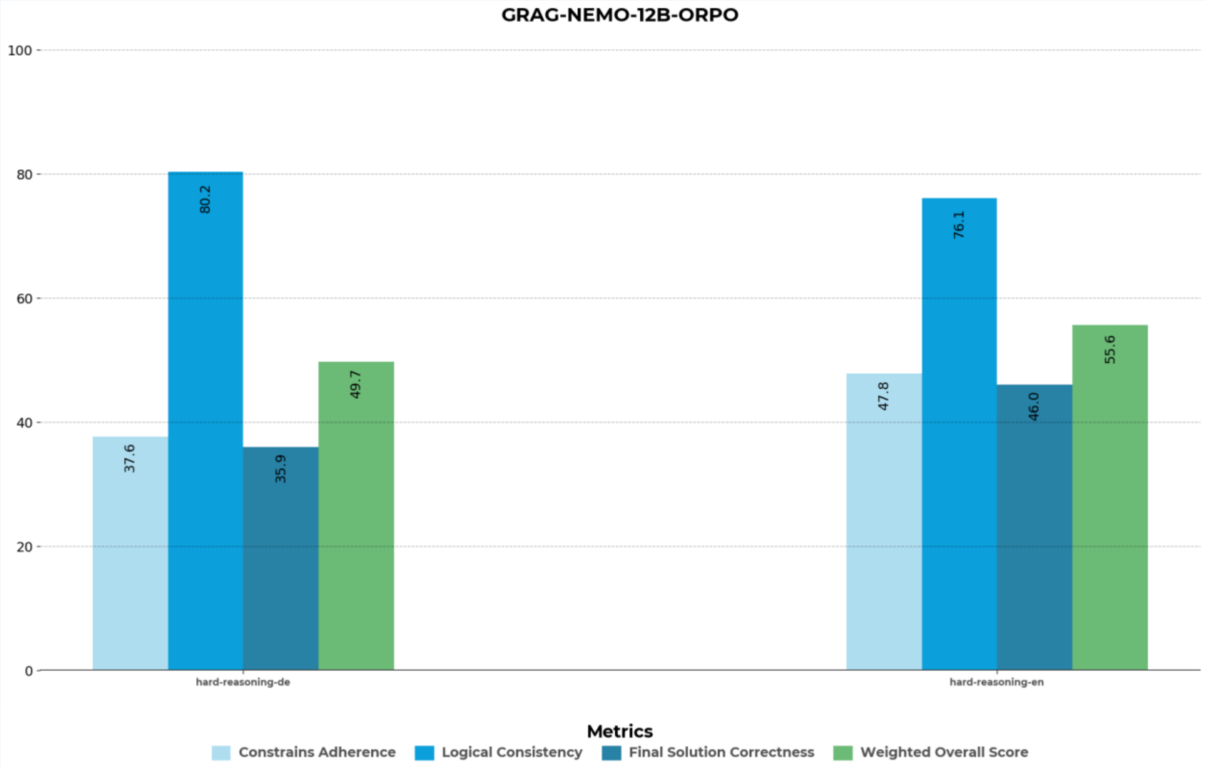
HARD-BENCHMARK | Evaluation of NEMO-12B



NEMO-12B-INSTRUCT, being our largest evaluated model at 12 billion parameters, demonstrates interesting reasoning capabilities in its instruction-tuned version. In German reasoning tasks, the model achieves a constraint adherence of 31.4% while showing particularly strong logical consistency at 74.3%. Its final solution correctness of 29.6% contributes to a weighted score of 43.6%. These results suggest that instruction tuning can effectively develop reasoning capabilities even in non-English languages at this scale.

When examining English reasoning tasks, the model demonstrates more robust performance across metrics. The constraint adherence improves to 44.6%, while logical consistency reaches an impressive 76.9%. With a final solution correctness of 44.4%, the model achieves a weighted score of 54.2%. This enhanced performance in English tasks aligns with patterns we've seen in other models, suggesting that instruction tuning at this scale can develop particularly strong English language reasoning capabilities.

What makes these results especially informative is how they demonstrate the relationship between model size and reasoning capabilities. At 12 billion parameters, NEMO shows particularly strong logical consistency scores in both languages, suggesting that larger model size can help develop more robust reasoning frameworks. However, the gap between logical consistency and other metrics like constraint adherence and solution correctness raises interesting questions about how instruction tuning at this scale might be optimized to achieve more balanced performance across all aspects of reasoning. This baseline performance provides valuable insights for understanding how different training approaches might leverage NEMO's larger capacity.



Our GRAG-NEMO-12B-ORPO model presents an intriguing alternative path to developing reasoning capabilities compared to Mistral's instruction tuning approach. In German reasoning tasks, our training methodology achieves constraint adherence of 37.6% and showcases strong logical consistency at 80.2%. With a final solution correctness of 35.9%, it reaches a weighted score of 49.7%. The higher logical consistency score is particularly interesting, suggesting that our targeted training approach may develop different aspects of reasoning compared to traditional instruction tuning.

For English reasoning tasks, both approaches demonstrate robust capabilities with their own distinct characteristics. Our model shows strong constraint adherence at 47.8% and maintains impressive logical consistency at 76.1%. With a final solution correctness of 46.0%, it achieves a weighted score of 55.6%. These results illustrate how different training philosophies can achieve similar overall outcomes while emphasizing different aspects of performance.

What makes this comparison particularly fascinating is how it illuminates different pathways to developing sophisticated reasoning capabilities in larger language models. While instruction tuning takes a comprehensive approach to developing general capabilities, our ORPO methodology focuses specifically on enhancing reasoning and output quality. The fact that both approaches achieve strong results through different means suggests that model development isn't limited to a single training philosophy. This insight could be especially valuable for researchers exploring ways to optimize larger models, as it demonstrates that targeted training strategies can be as effective as broader instruction tuning approaches. The similar overall performance levels, achieved through distinct methodologies, suggest that multiple viable paths exist for developing advanced reasoning capabilities in language models, each with its own unique strengths and characteristics.

HARD-BENCHMARK | CONCLUSION

The evaluations on this part of the HARD-BENCHMARK across various models highlight the nuanced interplay between model size, training methodologies, and performance in complex reasoning tasks. While larger models like NEMO-12B-INSTRUCT demonstrate superior baseline performance due to their parameter count and instruction tuning, the findings reveal that smaller models, such as PHI-4B and Mistral-7B, can achieve somewhat competitive reasoning capabilities to GPT-3.5-Turbo for underrepresented language with targeted training strategies like ORPO and SFT.

Key takeaways from the HARD-BENCHMARK Evaluation include:

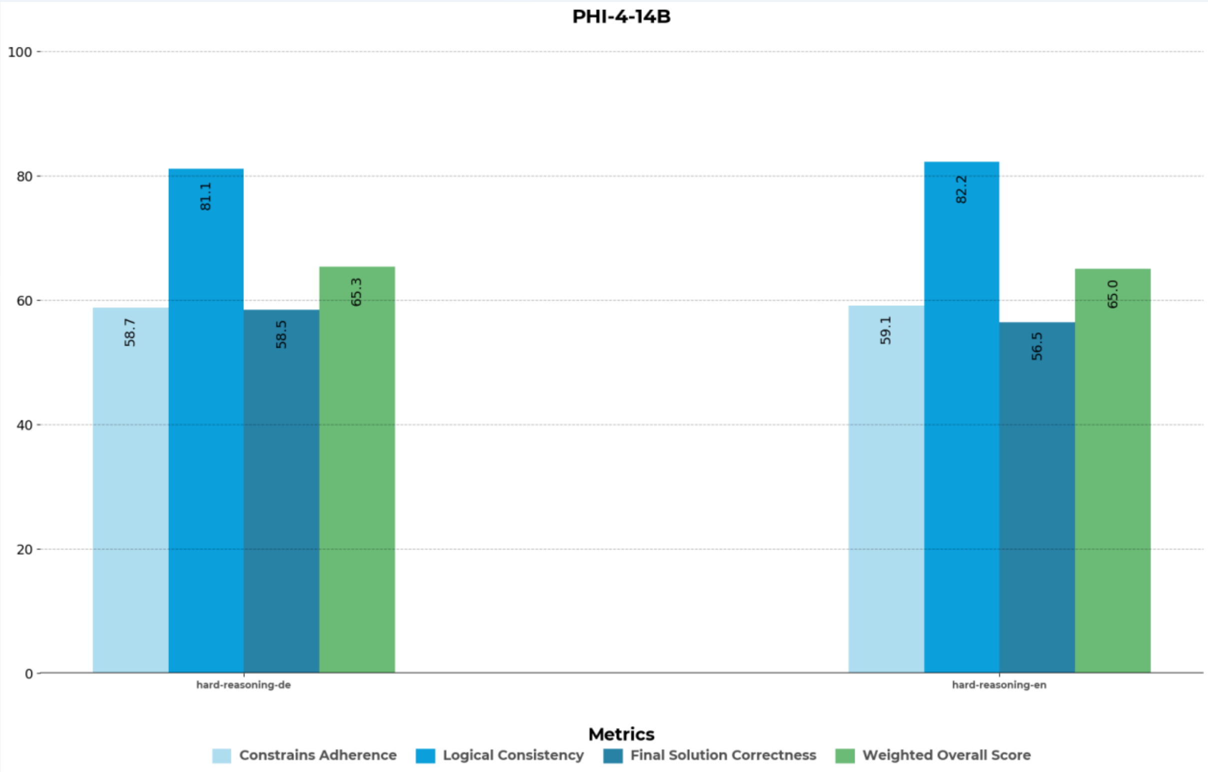
1. **Logical Consistency as a Strength:** Across all models, logical consistency emerged as a standout metric, with even compact models like PHI-4B and Mistral-7B achieving scores approaching those of larger commercial models like GPT-4o. This suggests that logical reasoning could be effectively cultivated through specialized training, regardless of model size.
2. **Training Impact on Performance:** The comparison of baseline and trained versions (e.g., ORPO and Merged models) highlights the significant role of targeted training in enhancing reasoning capabilities. Notably, the GRAG-LLAMA-8B and GRAG-NEMO-12B models achieved performance metrics comparable to much larger, commercially tuned models, underscoring the effectiveness of the GRAG training methodology.
3. **Challenges with Source Attribution:** A consistent challenge across all models was the handling of source citations in the hard-qa-with-multiple-references task. This limitation points to an area for future improvement, especially for applications requiring robust reference attribution. This is why we build these models to generate training data for such “hard-citation”-tasks at scale & locally.
4. **Language-Specific Variations:** The evaluations revealed language-specific performance dynamics. For instance, while models generally exhibited stronger logical consistency in English reasoning, German reasoning presented challenges in constraint adherence, even with additional training data.
5. **Scalability of Reasoning Capabilities:** The results emphasize that even smaller models can develop meaningful reasoning abilities. This finding is particularly relevant for scenarios where computational efficiency and task-specific fine-tuning are prioritized.

Overall, this evaluation illustrates the potential for smaller models to compete in the future with larger counterparts in specific domains through targeted training.

The strong logical consistency scores achieved by all trained models demonstrate that strategic training approaches can bridge the gap between open-source models and state-of-the-art commercial systems, paving the way for more efficient and capable language models. Future research should explore further optimization of training dynamics, especially in areas like source attribution and multilingual reasoning, to unlock the full potential of these models.

Looking back at the results from our experiment to continue pre-training, SFT- & ORPO-Training on the Instruct Version of PHI, we would assume that our training approach could have benefited if we would have directly used the Instruct-Versions of the Models instead of the Base-Models. Nevertheless, we are happy to achieve similar performance for reasoning-tasks matching or exceeding instruct versions from AI-Labs like Meta, Mistral & Microsoft. Our next steps will include to generate more training-samples for the hard-qa-multiple-references task with our GRAG-Models to even surpass performance in source citation of commercial models like GPT-4o with small open language models. This is what we think is the path to reliable, trustworthy & explainable AI-Systems, where the users can understand on which sources the answer is generated.

Open-Source Outlook for Reasoning



The comparison between PHI-4-14B and VANILLA-GPT-4o highlights the growing potential of open-source LLMs. Despite being significantly smaller, PHI-4-14B consistently outperforms its larger counterpart in key metrics like Logical Consistency and Weighted Overall Score. **Achieving over 65 in overall scores, PHI-4-14B demonstrates that smaller, optimized models can rival larger, resource-intensive commercial ones.**

This marks a shift in the open-source AI landscape, where efficiency and accessibility are taking precedence. Smaller models like PHI-4-14B reduce hardware and energy demands, making advanced AI more sustainable and widely available. As open-source communities innovate further, we can expect even greater performance from compact models, democratizing AI development and expanding its impact across industries.

GRAG Embedding Models

In this section, we present several embedding models, trained on the challenging GRAG-Embedding-Triples-Hessian-AI dataset. The dataset, comprising approximately 294.000 triple samples, was specifically designed to train the effectiveness of embedding models on nuanced query-response relationships.

The samples were synthetically generated using questions and answers derived from contextual chunks of the same Wikipedia pages. Each sample contained:

- A **positive chunk**: directly related to the query (question or answer).
- A **negative chunk**: the next best-matching chunk from the same Wikipedia page, determined using the BGE-M3 embedding model.

For instance, the following example illustrates the structure of a sample:

Query:

What are the special characteristics of the European mole's food stores before the winter months?

Positive Example:

Mainly before the winter months, the European mole stores earthworms in its tunnel system. Some such storage chambers can consist of up to 790 earthworms with a total weight of 1.5 kg, which is enough food for more than six weeks. ...

Negative Example:

The European mole's main diet consists of earthworms and their cocoons. In various studies in Scotland and southern Poland, evidence of earthworms was found in 85 to 100 % of the stomach remains analyzed. ...

This challenging structure ensures that the models must accurately capture subtle differences between closely related yet distinct contexts.

We trained the following models using the dataset:

GRAG-BGE-M3-TRIPLES-HESSIAN-AI

GRAG-BGE-M3-TRIPLES-MERGED-HESSIAN-AI

GRAG-BGE-M3-MERGED-x-SNOWFLAKE-ARCTIC-HESSIAN-AI

GRAG-UAE-LARGE-V1-TRIPLES-HESSIAN-AI

GRAG-UAE-LARGE-V1-TRIPLES-MERGED-HESSIAN-AI

Training and Merging Process

1. **Initial Training:** Both the BGE-M3 and UAE-Large-V1 base models were trained on the dataset.

2. **Model Merging:** Post-training, each trained model was merged back with its respective base model to maintain prior performances in other languages and tasks.

3. **Further Re inement:** For BGE-M3, an additional merging step was conducted. The trained and merged BGE-M3 model was further merged with the fine-tuned version of BGE-M3 provided by Snowflake (Snowflake/snowflake-arctic-embed-l-v2.0).

The results highlight the success of our training and merging strategies, particularly in improving model accuracy on such a challenging dataset. This section serves as a foundation for the analysis and comparison of model performances presented in the subsequent sections.

Evaluation on MTEB-Tasks (German Subsets)

Classification

- AmazonCounterfactualClassification
- AmazonReviewsClassification
- MassiveIntentClassification
- MassiveScenarioClassification
- MTOPDomainClassification
- MTOPIntentClassification

Pair Classification

- FalseFriendsGermanEnglish
- PawsXPairClassification

Retrieval

- GermanQuAD-Retrieval
- GermanDPR

STS (Semantic Textual Similarity)

- GermanSTSBenchmark

Evaluation on GRAG-EMBEDDING-BENCHMARK

The Embedding-Models receive a query and a context array. Each chunk of the context array and the query gets embedded and then the similarity gets calculated between each chunk and the query. If the relevant_context provided is the chunk with the highest similarity it is scored with 1 otherwise with 0.

query string · *lengths* 180⊖220 16.3%	relevant_context string · *lengths* 340⊖579 34.7%	context_array sequence · *lengths* 4⊖5 30.8%
Reine Datenklassen ohne fachliche Funktionalität gelten im Domain-driven Design als Code-Smell, da sie ein Domänenmodell ohne Fachlogik darstellen, was als Antipattern angesehen wird.	Die Klassen des Domänenmodells enthalten im Domain-driven Design sowohl die Daten als auch die gesamte Funktionalität der umzusetzenden Fachlichkeit, also die gesamte Fachlogik. Reine Datenklassen nur mit Zugriffsmethoden aber ohne fachliche Funktionalität gelten als Code-Smell. Ein auf Datenklassen aufbauendes Domänenmodell wird anämisch genannt und gilt demzufolge als Antipattern, da es ein Domänenmodell ohne Fachlogik beschreibt.	["Domain-driven Design ist an keinen bestimmten Softwareentwicklungsprozess gebunden, orientiert sich aber an agiler Softwareentwicklung. Insbesondere setzt es iterative Softwareentwicklung und eine enge Zusammenarbeit zwischen Entwicklern und Fachexperten voraus.", "Der Sinn jeder Software ist es, die Aufgabenstellungen einer bestimmten Anwendungsdomäne zu unterstützen. Um dies erfolgreich leisten zu können, muss die Software harmonisch zu der Fachlichkeit der Anwendungsdomäne passen, für die sie bestimmt ist. Domain-driven Design ermöglicht dies, indem die Software grundlegende Konzepte und Elemente der Anwendungsdomäne sowie deren Beziehungen modelliert.", "Die Architektur ist geprägt durch die Existenz einer expliziten Geschäftslogikschicht. Diese Schicht soll die Domänen-Klassen von anderen Funktionen des Systems entkoppeln und möglichst leicht erkennbar machen. Verschiedene Architekturstile können eingesetzt werden, um die Geschäftslogikschicht einzubetten. Dazu zählen die Schichtenarchitektur und die hexagonale Architektur.", "Die Klassen des Domänenmodells enthalten im Domain-driven Design sowohl die Daten als auch die gesamte Funktionalität der umzusetzenden Fachlichkeit, also die gesamte Fachlogik. Reine Datenklassen nur mit Zugriffsmethoden aber ohne fachliche Funktionalität gelten als Code-Smell. Ein auf Datenklassen aufbauendes Domänenmodell wird anämisch genannt und gilt demzufolge als Antipattern, da es ein Domänenmodell ohne Fachlogik beschreibt."]

GRAG BGE-M3 Models

Comparison between Base-Model (BGE-M3), Finetuned Model (GRAG-BGE) and Merged Model with Base-Model (Merged-BGE)

TASK	BGE	GRAG-BGE	Merged-BGE	GRAG vs. BGE	Merged vs. BGE
AmazonCounterfactualClassification	0,6908	0,5449	**0,7111**	-14,59%	**2,03%**
AmazonReviewsClassification	**0,4634**	0,2745	0,4571	-18,89%	**-0,63%**
FalseFriendsGermanEnglish	**0,5343**	0,4777	0,5338	-5,67%	**-0,05%**
GermanQuAD-Retrieval	**0,9444**	0,8714	0,9311	-7,30%	**-1,33%**
GermanSTSBenchmark	0,8079	0,7921	**0,8218**	-1,58%	**1,39%**
MassiveIntentClassification	**0,6575**	0,4884	0,6522	-16,90%	**-0,52%**
MassiveScenarioClassification	0,7355	0,5837	**0,7381**	-15,19%	**0,25%**
GermanDPR	**0,8265**	0,7210	0,8159	-10,54%	**-1,06%**
MTOPDomainClassification	0,9121	0,7450	**0,9139**	-16,71%	**0,17%**
MTOPIntentClassification	**0,6808**	0,4516	0,6684	-22,92%	**-1,25%**
PawsXPairClassification	**0,5678**	0,5077	0,5710	-6,01%	**0,33%**

We successfully merged the Snowflake/snowflake-arctic-embed-l-v2.0 model with our merged version of the BGE-M3 model by performing distillation. This process ensured that both models had the same dimensions, enabling a seamless and efficient merging process.

Comparison between Snowflake-Model (Arctic based on BGE-M3), Merged Model with Base-Model (Merged-BGE) and our Merged-Model merged with Snowflake/snowflake-arctic-embed-l-v2.0

TASK	Snowflake Arctic	Merged-BGE	Merged-Snowflake	GRAG vs. Snowflake	Merged-Snowflake vs. Snowflake	Merged-Snowflake vs. Merged-BGE
AmazonCounterfactualClassification	0,6587	0,7111	**0,7152**	5,24%	5,65%	**0,41%**
AmazonReviewsClassification	0,3697	0,4571	0,4577	8,74%	8,80%	**0,06%**
FalseFriendsGermanEnglish	0,5360	0,5338	**0,5378**	-0,22%	0,18%	**0,40%**
GermanQuAD-Retrieval	0,9423	0,9311	**0,9456**	-1,12%	0,33%	**1,45%**
GermanSTSBenchmark	0,7499	0,8218	**0,8558**	7,19%	10,59%	**3,40%**
MassiveIntentClassification	0,6778	0,6522	**0,6826**	-2,56%	0,48%	**3,04%**
MassiveScenarioClassification	0,7375	0,7381	**0,7494**	0,06%	1,19%	**1,13%**
GermanDPR	0,8367	0,8159	**0,8330**	-2,08%	-0,37%	**1,71%**
MTOPDomainClassification	0,9080	0,9139	**0,9259**	0,59%	1,79%	**1,20%**
MTOPIntentClassification	0,6675	0,6684	**0,7143**	0,09%	4,68%	**4,59%**
PawsXPairClassification	0,5887	0,5710	**0,5803**	-1,77%	-0,84%	**0,93%**

GRAG-EMBEDDING-BENCHMARK:

Model Name	Accuracy
bge-m3 (base-model)	0.8806
GRAG-BGE-M3-TRIPLES-HESSIAN-AI	0.8857
GRAG-BGE-M3-TRIPLES-MERGED-HESSIAN-AI	**0.8866**
GRAG-BGE-M3-MERGED-x-SNOWFLAKE-ARCTIC-HESSIAN-AI	**0.8866**

GRAG UAE-Large-V1 Models

Comparison between Base-Model (UAE), Finetuned Model (GRAG-UAE) and Merged Model with Base-Model (Merged-UAE)

TASK	UAE	GRAG-UAE	Merged-UAE	GRAG vs. UAE	Merged vs. UAE
AmazonCounterfactualClassification	**0,5650**	0,5449	0,5401	-2,01%	-2,48%
AmazonReviewsClassification	0,2738	0,2745	**0,2782**	0,08%	0,44%
FalseFriendsGermanEnglish	**0,4808**	0,4777	0,4703	-0.32%	-1,05%
GermanQuAD-Retrieval	0,7811	0,8353	**0,8628**	5,42%	8,18%
GermanSTSBenchmark	0,6421	0,6568	**0,6754**	1,47%	3,33%
MassiveIntentClassification	**0,5139**	0,4884	0,4714	-2,55%	-4,25%
MassiveScenarioClassification	0,6062	0,5837	**0,6111**	-2,25%	0,49%
GermanDPR	0,6750	0,7210	**0,7507**	4,60%	7,57%
MTOPDomainClassification	0,7625	0,7450	**0,7686**	-1,75%	0,61%
MTOPIntentClassification	**0,4994**	0,4516	0,4413	-4,77%	-5,80%
PawsXPairClassification	**0,5452**	0,5077	0,5162	-3,76%	-2,90%

GRAG-EMBEDDING-BENCHMARK:

Model Name	Accuracy
UAE-Large-V1 (base-model)	0.8393
GRAG-UAE-LARGE-V1-TRIPLES-HESSIAN-AI	0.8763
GRAG-UAE-LARGE-V1-TRIPLES-MERGED-HESSIAN-AI	**0.8771**

GRAG Whisper Model (Speech-to-Text)

The GRAG-WHISPER-LARGE-v3-TURBO model is a fine-tuned variant of Whisper Large v3 Turbo, optimized for high-accuracy speech-to-text transcription. This model was trained on a meticulously curated 13-hour dataset featuring spoken German mixed with English business phrases, selected to enhance recognition performance in multilingual and domain-specific contexts.

Training Data

The training data for this model comprises conversational spoken German interspersed with English business phrases. The data was meticulously selected and processed to ensure optimal recognition performance, particularly in contexts where code-switching or domain-specific language is common.

To address potential ethical concerns related to voice cloning, the dataset will not be published. Its usage rights are explicitly limited to training speech-to-text models, ensuring compliance with ethical and legal guidelines.

Evaluations – Word Error Rate (WER) | avemio/ASR-GERMAN-MIXED-TEST

The model demonstrates significant improvements in word error rates (WER) across various test datasets when compared to the base **openai/whisper-large-v3-turbo** and **primeline/whisper-large-v3-turbo-german** models:

Test Dataset	openai-whisper-large-v3-turbo	GRAG-WHISPER-LARGE-v3-TURBO	primeline-whisper-large-v3-turbo-german
Tuda-De	8.195	**6.360**	6.441
common_voice_19_0	3.839	3.249	**3.217**
Multilingual Librispeech	3.202	2.071	**2.067**
All (Combined)	3.641	2.633	**2.630**

The GRAG-WHISPER-LARGE-v3-TURBO model consistently outperforms the base Whisper model across all datasets, showcasing its robustness in diverse multilingual and domain-specific scenarios. Notably, the WER on the Tuda-De dataset improved by nearly 22%, highlighting its enhanced performance for German speech-to-text tasks.

Conclusion

The GRAG-WHISPER-LARGE-v3-TURBO model exemplifies the potential of fine-tuned speech-to-text models tailored for German business use cases like transcribing meetings. Its superior performance across benchmarks demonstrates the efficacy of the targeted training process and positions this model as a valuable asset for transcription tasks involving German and mixed-language speech.

Conclusion:

The GRAG-Suite - Advancing Open German AI Research

Our comprehensive evaluation of the GRAG Model Suite demonstrates significant progress in developing efficient, open-source models specialized for German language tasks. The suite's performance across LLMs, Embedding Models, and Speech Recognition highlights three key achievements:

Efficient Model Scaling

- Successfully trained 4B-12B parameter models achieving competitive performance with much larger commercial models
- Demonstrated that targeted training can effectively bridge capability gaps, particularly in reasoning and source attribution
- Established baseline performance metrics for German Language Model evaluation

Technical Innovations

- **Data:** Providing all Building Blocks (Models) to synthetically generate Trainingdata locally with Open-Source Models on your own Data
- **LLMs:** Developed effective training methodology combining CPT, SFT, and ORPO approaches and showcase Merging-Benefits for Llama-Architecture
- **Embeddings:** Created specialized models outperforming base versions in German tasks while maintaining multilingual capabilities
- **Speech-to-Text:** Achieved state-of-the-art WER rates for German mixed-language transcription comparable to the best German Speech-to-Text Model from **primeline**

Future Research Directions

- Generate high-quality training data using current GRAG models to improve source attribution capabilities
- Develop specialized evaluation benchmarks for German language tasks

The GRAG Model Suite establishes a foundation for collaborative German AI research, demonstrating that focused, open-source development can create efficient, capable models for specific language domains. We invite the research community to build upon these results, particularly in:

- Optimizing training approaches for smaller models
- Developing German-specific benchmarks and evaluation metrics
- Expanding multilingual capabilities while maintaining computational efficiency

This work represents a significant step toward democratizing AI development for German language applications, proving that targeted research and open collaboration can advance state-of-the-art performance in specialized domains.

You can now access all Models & datasets **here**.

Wish you all the best while experimenting!

The GRAG-Team

Marcel Rosiak, Soumya Paul, Siavash Mollaebrahim, Zain Ul Haq

DISCLAIMER

This study was conducted as part of an effort to train and evaluate artificial intelligence (AI) systems. The data collected, processed, and analyzed during this study was handled in accordance with applicable ethical guidelines and data protection regulations.

The findings and conclusions presented in this study are based on the specific methodologies, datasets, and models employed, which may have inherent limitations. The outcomes of the study should not be interpreted as universally applicable or as guarantees of AI system performance under conditions different from those tested.

Avemio AG provides no warranty (either express or implied) for the accuracy, completeness, and timeliness of the information and opinions published in this training study. In the event of unintentional biases or errors in the data, model, or conclusions, the research team welcomes feedback and will take appropriate corrective measures.

The authors of this study are solely responsible for its content. The views expressed do not necessarily reflect the positions of the affiliated institutions, sponsors, or stakeholders.

PUBLISHER

Avemio AG
Peter-Sander-Str. 41c
55252 Mainz-Kastel
www.avemio.com

EDITORIAL & DESIGN

avemio.digital GmbH
Marcel Rosiak
Schatzbogen 50-52
81829 München
E-Mail: grag@avemio.digital
https://avemio.digital

IN COOPERATION WITH

Hessian Center for Artificial Intelligence hessian.AI
represented by Darmstadt University of Technology
Karolinenplatz 5
64289 Darmstadt
E-Mail: communications@hessian.ai
https://hessian.ai/