

ECG-LLM– training and evaluation of domain-specific large language models for electrocardiography

Lara Ahrens¹, Wilhelm Haverkamp², Nils Strodthoff^{1*}

¹AI4Health Division, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany.

²Department of Cardiology, Angiology and Intensive Care Medicine, Charité Campus Mitte, German Heart Center of the Charité-University Medicine Berlin, Berlin, Germany.

*Corresponding author(s). E-mail(s): nils.strodthoff@uol.de;
Contributing authors: lara.ahrens@uol.de;
wilhelm.haverkamp@dhzc.charite.de;

Abstract

Domain-adapted open-weight large language models (LLMs) offer promising healthcare applications, from queryable knowledge bases to multimodal assistants, with the crucial advantage of local deployment for privacy preservation. However, optimal adaptation strategies, evaluation methodologies, and performance relative to general-purpose LLMs remain poorly characterized. We investigated these questions in electrocardiography, an important area of cardiovascular medicine, by finetuning open-weight models on domain-specific literature and implementing a multi-layered evaluation framework comparing finetuned models, retrieval-augmented generation (RAG), and Claude Sonnet 3.7 as a representative general-purpose model. Finetuned Llama 3.1 70B achieved superior performance on multiple-choice evaluations and automatic text metrics, ranking second to Claude 3.7 in LLM-as-a-judge assessments. Human expert evaluation favored Claude 3.7 and RAG approaches for complex queries. Finetuned models significantly outperformed their base counterparts across nearly all evaluation modes. Our findings reveal substantial performance heterogeneity across evaluation methodologies, underscoring assessment complexity. Nevertheless, domain-specific adaptation through finetuning and RAG achieves competitive performance with proprietary models, supporting the viability of privacy-preserving, locally deployable clinical solutions.

Keywords: Natural Language Processing, Large Language Models, Domain Specialization, Cardiology, Electrocardiography

1 Introduction

Large-language models (LLMs) used as question-answering models [1] represent a promising approach for clinicians to cope with the ever-increasing amount of medical knowledge. If properly validated and implemented, such systems could help reduce misdiagnoses and associated treatment errors and could support the identification of optimal, evidence-based treatment strategies. Such a paradigm could be implemented either through commercial general-purpose LLMs, with associated privacy risks, or through smaller domain-adapted open-weight models. In this study, we focus on cardiology with an emphasis on electrocardiography and set out to detail practical techniques for domain specialization of LLMs and compare them thoroughly in a comprehensive evaluation approach; see Figure 1 for a schematic overview.

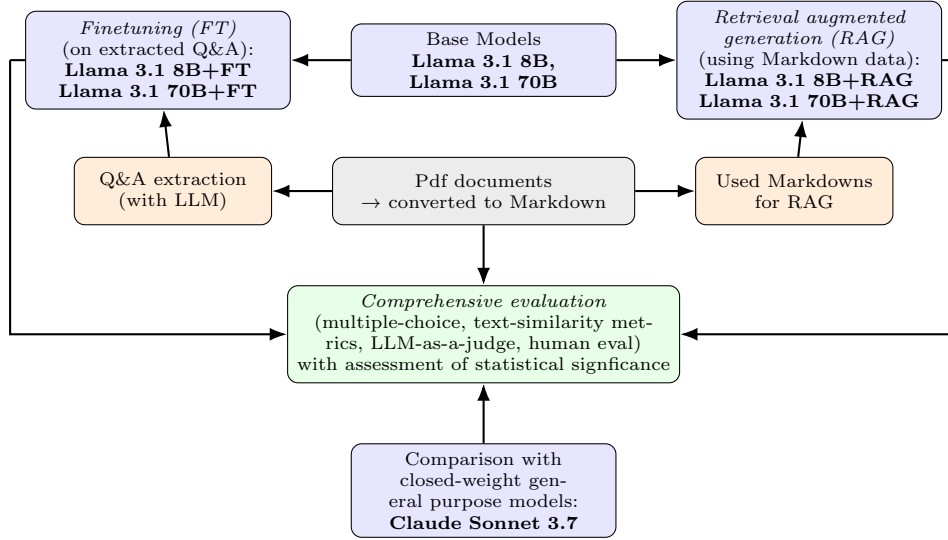


Fig. 1 Schematic overview of the core methodology of this study: We study finetuning and retrieval augmented generation as two possible paths towards domain specialization of Llama 3 open-weight LLMs in the domain of electrocardiography. Model performance is assessed in a comprehensive evaluation procedure covering four different categories comparing domain-specialized models in comparison to the respective base models and a commercial general-purpose LLM.

For natural language processing (NLP) tasks such as question answering, large language models (LLMs) are becoming increasingly important. Models with varying capabilities, sizes, and knowledge bases are currently being developed [2]. However,

LLMs tend to hallucinate when knowledge, in particular domain-specific knowledge, is missing, generating incorrect or fabricated information [3], which represents a major hurdle in safety-critical domains such as medicine. The most powerful LLMs, such as GPT-4o and Claude Sonnet 3.7, are typically closed-source models, which prohibits local hosting and limits domain knowledge adaptation. In the medical domain, additional challenges include specialized knowledge, specialized vocabulary, reliability, and data protection [4]. Consequently, large closed-source models may not be ideal for medical tasks. Therefore, we explore smaller open-weight LLMs that can be locally hosted and adapted to specific medical contexts.

However, the superiority of domain-specialized language models over their general-purpose counterparts should not be taken for granted. Prior work [5] found no evidence for significant improvements when using appropriate prompting, and the same holds for finetuned commercial LLMs [6]. Nevertheless, finetuned domain-specialized models have been explored in many domains, including medical subdomains such as radiology [7], material science [8], and mathematics [9].

Identifying the most effective methods for domain specialization of pretrained general-purpose LLMs remains an active area of research. Possible methods range from continual pretraining [10] and supervised finetuning [11–13] to retrieval-augmented generation (RAG) [14–16]. Whereas finetuning was previously believed to primarily adjust output style, recent work has demonstrated its potential for knowledge injection through question-answer pairs [11]. This aligns with previous research showing that finetuning can fundamentally improve the model’s domain knowledge [4, 17–20]. RAG represents a complementary approach that enriches the model’s responses with external, factual information [14]. During RAG, the LLM is supplemented with retrieved, contextually relevant data to generate fact-based and verifiable answers [15].

Comprehensive reviews highlight the growing potential of LLMs in cardiovascular medicine, including applications in patient education, clinical decision support, and workflow automation [21]. Studies have demonstrated that finetuning LLMs on small-scale cardiology datasets can substantially improve clinical text classification and information extraction [22], and LLMs have shown promise in automating discharge summary generation for cardiac patients [23]. In evaluation studies, Lee et al. benchmarked general-purpose LLMs on cardiology case questions modeled after the American College of Physicians’ MKSAP examination, finding that GPT-4 achieved performance comparable to seasoned cardiologists [24]. While these studies demonstrate the potential of LLMs in cardiology, they either focus on narrow application domains or do not investigate the effects of domain specialization through comprehensive evaluation. A recent study [20] demonstrated that a finetuned Llama 2 outperformed its original base model (the starting point before domain adaptation) across all automated metrics and human evaluations. Our work builds on state-of-the-art base models from the Llama 3 family [25] and provides a more comprehensive evaluation. Regarding RAG, a recent cardiology study showed that RAG can improve LLMs’ knowledge of ECG and outperform OpenAI models across different tasks [26]. This RAG evaluation is closely related to our work, which additionally provides a comparative assessment of finetuning and a more comprehensive evaluation comprising different methods.

Recent works have extended the scope to multimodal applications that leverage LLMs as a core component, including multimodal multi-agent models [27, 28], conversational ECG interpretation models [29], and multimodal medical text retrieval models [30]. Since most multimodal assistant models leverage general-purpose LLMs, they could potentially benefit from domain-adapted LLMs as explored in this work.

2 Results

We finetune Llama 3.1 8B and 70B open-weight (instruct) models on question-answer pairs extracted from domain-specific literature by prompting Llama 3.3 70B. We use the same references as a basis for retrieval-augmented generation and evaluate models using multiple-choice questions extracted from the same corpus, automatic text similarity metrics, LLM-as-a-judge evaluation, and human expert evaluation by an experienced cardiologist. We employ empirical bootstrapping to derive statistically robust rankings.

2.1 Multiple-choice evaluation

For the multiple-choice evaluation, we assess three dataset subsets: “full” contains the entire dataset excluding training data (27,774 samples), “special” contains questions from documents highlighted by the human expert as particularly relevant (1,219 samples), and “checked” contains multiple-choice questions verified by a human expert with long-standing experience in electrocardiography and cardiac electrophysiology (534 samples). Table 1 presents the results across these three tests. The finetuned versions and RAG-enhanced models consistently outperform general-purpose models, with the two domain-specialized 70B models performing best. Notably, although model accuracies vary slightly, the ranking remains consistent across all three subsets.

Table 1 Performance of different models across multiple-choice evaluation sets. Values represent accuracy percentages with rank in parentheses. Domain-adapted models explored in this work are highlighted in bold face. Domain-adapted models dominate over general-purpose models in this setting. Special = specialized cardiovascular/ECG questions; Full = complete question set; Checked = manually validated subset; RAG = retrieval-augmented generation; FT = finetuning.

Model	Special	Full	Checked
Llama 3.1 70B + FT	90.2 (1)	92.0 (1)	88.2 (1)
Llama 3.1 70B + RAG	87.9 (2)	90.2 (2)	86.5 (2)
Llama 3.1 8B + RAG	86.1 (3)	88.5 (3)	85.0 (3)
Llama 3.1 8B + FT	84.9 (3)	87.5 (6)	84.3 (4)
Claude Sonnet 3.7	82.3 (5)	88.0 (4)	81.7 (5)
Llama 3.1 70B	82.4 (6)	88.2 (5)	81.5 (6)
Llama 3.1 8B	73.3 (7)	81.6 (7)	73.0 (7)

2.2 Text similarity metrics

For automatic evaluation, we assessed free text responses using BLEU [31], ROUGE-1, ROUGE-2, ROUGE-L [32], and BERTScore [33], where BLEU and ROUGE assess surface similarity and BERTScore assesses semantic similarity between generated response and the reference answer. Table 2 shows the F1 scores for these metrics; recall and precision values are provided in the Appendix E. The finetuned models consistently outperform state-of-the-art models, while RAG-enhanced models rank last, performing worse than the base models. Again, the ranking remains consistent across the different metrics, which assess different levels of text similarity.

Table 2 Text summarization performance metrics across different models. Values represent F1 scores (ROUGE, BERTScore) or BLEU scores with rank in parentheses. Domain-adapted models explored in this work are highlighted in bold face. Finetuned models excel in this category. RAG = retrieval-augmented generation; FT = finetuning.

Model	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	BLEU	BERTScore F1
Llama 3.1 70B + FT	0.4270 (1)	0.2449 (1)	0.3764 (1)	0.1289 (1)	0.3904 (1)
Llama 3.1 8B + FT	0.4063 (2)	0.2279 (2)	0.3568 (2)	0.1195 (2)	0.3700 (2)
Llama 3.1 70B	0.3852 (3)	0.1981 (3)	0.3208 (3)	0.0962 (3)	0.3476 (3)
Claude Sonnet 3.7	0.3661 (4)	0.1685 (5)	0.2992 (4)	0.0694 (6)	0.3395 (4)
Llama 3.1 8B	0.3537 (5)	0.1730 (4)	0.2906 (5)	0.0788 (4)	0.2936 (5)
Llama 3.1 8B + RAG	0.3481 (6)	0.1689 (5)	0.2831 (6)	0.0763 (5)	0.2861 (6)
Llama 3.1 70B + RAG	0.2557 (7)	0.1490 (7)	0.2183 (7)	0.0622 (7)	0.0181 (7)

2.3 LLM-as-a-judge

For the LLM-as-a-judge evaluation, a subset of the question-and-answer dataset was checked manually and corrected by a human expert to ensure quality. From this subset, we selected questions not part of the training dataset, resulting in 417 questions for evaluation. Figure 2 shows the results. Claude achieves the best results, while the finetuned 8B model shows substantial improvement, answering approximately 50 more questions correctly than its base model Llama 3.1 8B. The finetuned 70B model also improves, while other models remain stable or decrease in performance. Llama 3.1 8B and 70B with RAG achieve results similar to the finetuned versions. This metric shows statistically significant improvements in the finetuned models compared to their respective base models.

To verify the reliability of the LLM-as-a-judge assessment, a human expert evaluated a random subset of answers from the finetuned models labeled as correct or incorrect. The results are compiled in Appendix D. Overall, 80% of evaluations showed agreement between the LLM and human expert. Notably, all answers judged correct by the LLM were confirmed correct, but nearly 50% of answers judged incorrect by the LLM were considered correct by the human expert.

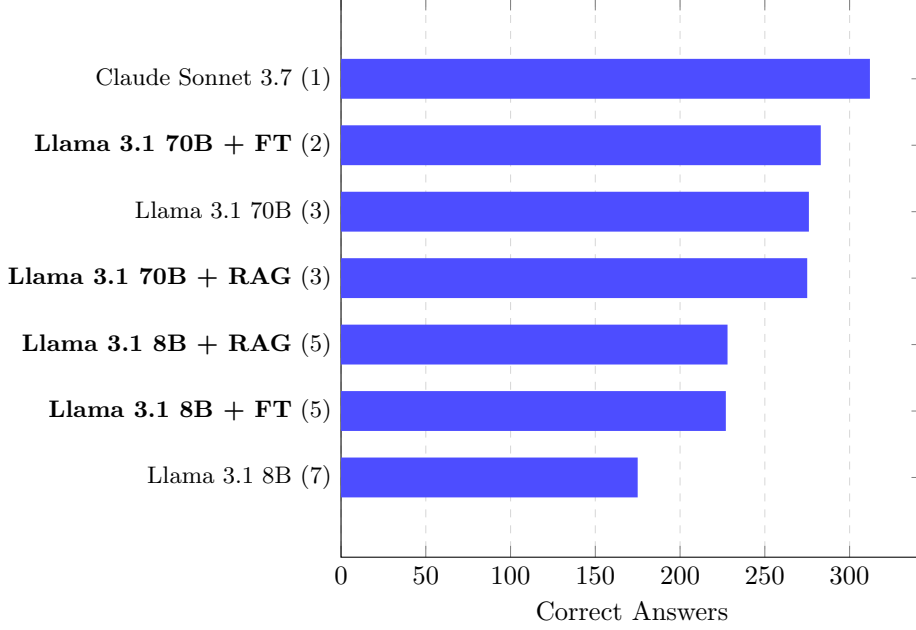


Fig. 2 Correct answers resulting from 417 questions evaluated with LLM-as-a-judge. The number in parentheses after the model name indicates the rank in the statistically robust model ranking. Domain-adapted models explored in this work are highlighted in bold face. Claude Sonnet 3.7 performs best, closely followed by Llama 3.1 70B models. RAG = retrieval-augmented generation; FT = finetuning.

2.4 Human evaluation

For the human evaluation, the expert provided ten questions asking for facts similar to those used for finetuning, as well as 40 more complex questions investigating relationships and differing in format and context. These complex questions, for example, ask for listings or descriptions of relationships between different characteristics. The expert then labeled the answers as incorrect, partially incorrect, correct but incomplete, or correct. The questions, model responses and expert evaluations are provided in the associated code repository.

Figure 3 shows the results of the human evaluation for the ten factual questions. The finetuned versions perform similarly to or better than their corresponding base models, providing more correct and correct-but-incomplete answers in total and never answering questions completely incorrectly. However, no model outperformed Claude Sonnet 3.7 and Llama 3.1 8B with RAG, which both answered all questions correctly. Llama 3.1 70B with RAG achieved one incomplete answer.

Figure 4 shows the results for the semantically more complex questions. Notably, the finetuned versions demonstrate slightly decreased performance compared to their base models. The number of completely incorrect answers increases for each finetuned model, resulting in rankings similar to Llama 3.1 8B. In contrast, Llama 3.1 8B with RAG outperforms its base model and reaches the performance of Llama 3.1 70B.

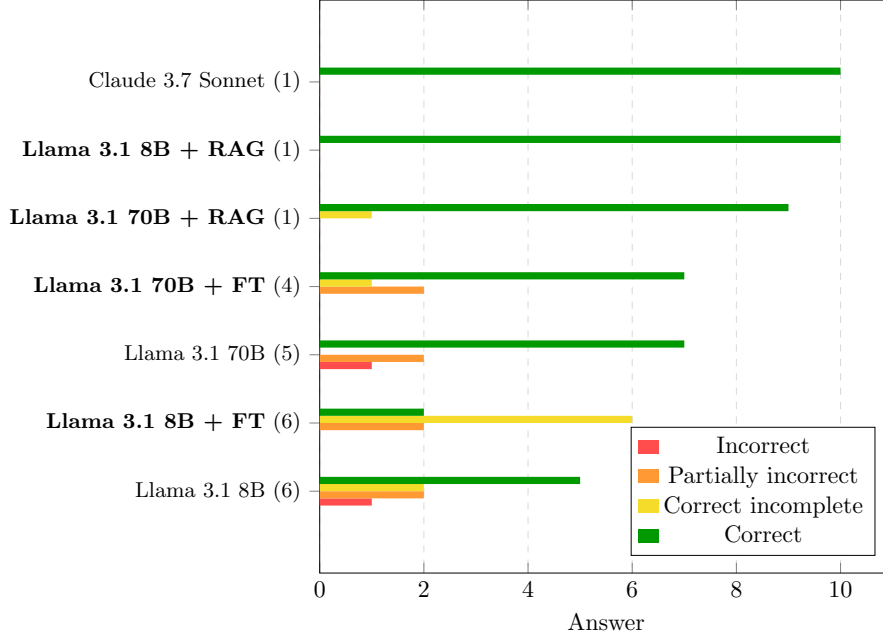


Fig. 3 Human evaluation of the ten factual questions. The number in parentheses after the model name indicates the rank in the statistically robust model ranking. Domain-adapted models explored in this work are highlighted in bold face. Both RAG models perform on par with the top-performing Claude 3.7 Sonnet. RAG = retrieval-augmented generation; FT = finetuning.

Unlike other evaluations, the 40-question assessment shows finetuned models either underperforming or achieving similar results to their corresponding base models.

To analyze the differences between the training data and the 40 complex human-provided questions, we calculated the average Flesch reading ease score. This metric measures text difficulty by analyzing word-to-sentence and syllable-to-word ratios, yielding a score between 0 and 100, where lower values indicate higher difficulty [34]. The human-provided questions averaged a Flesch score of 21.9, while the training data averaged 30.3, confirming that the human questions are semantically more complex.

2.5 Overall ranking

Table 3 shows the median ranks of each model across all evaluation methods. The finetuned Llama 3.1 70B demonstrates the strongest overall performance, followed by Claude Sonnet 3.7, which underperforms in multiple-choice and text-similarity evaluations. RAG is effective in some categories (LLM-as-a-judge and human evaluation) but underperforms in others and fails to consistently outperform the corresponding base models. Larger models (70B) consistently outperform smaller models (8B).

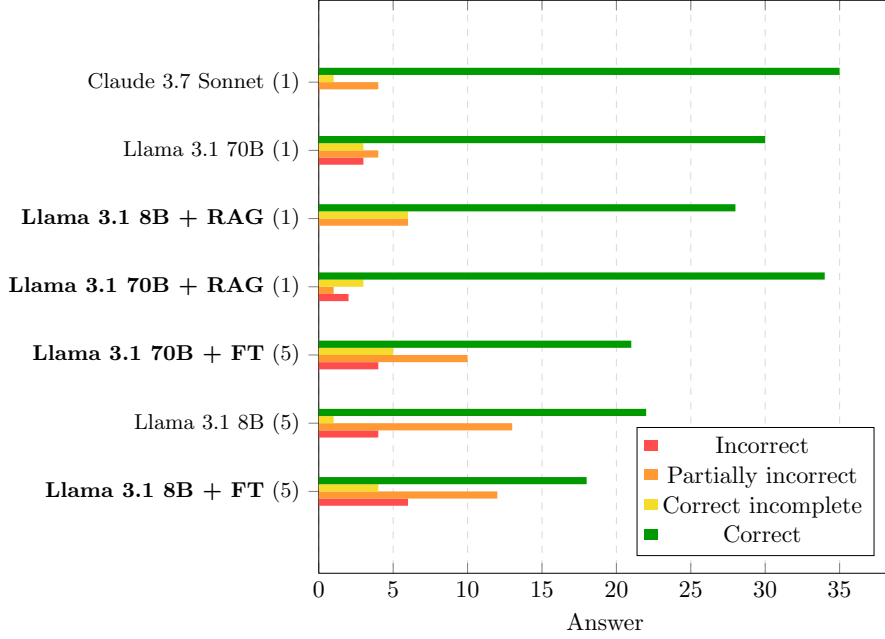


Fig. 4 Human evaluation of the 40 semantically complex questions. The number in parentheses after the model name indicates the rank in the statistically robust model ranking. Both RAG models as well as the Llama 3.1 70B base model perform on par with the best-performing Claude 3.7 Sonnet. Domain-adapted models explored in this work are highlighted in bold face. RAG = retrieval-augmented generation; FT = finetuning.

3 Discussion

3.1 Evaluation modes

Because the multiple-choice questions are LLM-generated, they may lack variation, and incorrect answer options might be obvious or occasionally correct. Nevertheless, consistent model rankings across all subsets, including the human-checked subset, suggest reliable results. Notably, the finetuned models outperform all state-of-the-art models on the checked and special subsets, indicating successful knowledge injection during finetuning. However, on the full multiple-choice set, the finetuned 8B model underperforms Claude and Llama 3.1 70B. The varying improvement margins also indicate that models with stronger baselines benefit less from finetuning.

Notably, all baseline models show poor performance in automatic text similarity evaluations. Since these metrics assess syntactic and semantic similarity [31–33], they demonstrate that finetuned responses align with reference answers, confirming knowledge acquisition. However, they cannot prove that state-of-the-art models underperform, as alternative correct formulations would score poorly despite being accurate. While these metrics effectively evaluate stylistic consistency of finetuned responses [11], they are less reliable for baseline models using alternative formulations. Evaluation based solely on BLEU and ROUGE scores can reveal style injection but cannot confirm

Table 3 Ranks for each of the models across the different evaluation categories (reporting medians in case of multiple results per category). Models are listed according to median rank across all categories. RAG = retrieval-augmented generation; FT = finetuning.

Model	Multiple-Choice	Text Similarity Metrics	LLM-as-a-Judge	Human Evaluation	Median Rank
Llama 3.1 70B + FT	1	1	2	3	1.5
Claude Sonnet 3.7	5	4	1	1	2.5
Llama 3.1 70B + RAG	3	7	3	1	3
Llama 3.1 70B	6	3	3	3	3
Llama 3.1 8B + RAG	2	6	5	1	3.5
Llama 3.1 8B + FT	4	2	5	5.5	4.5
Llama 3.1 8B	7	5	7	5.5	6

superior knowledge. However, the agreement across all metrics, including BERTScore’s higher-level semantic alignment, supports the robustness of these findings.

The LLM-as-a-judge also ranks finetuned models higher than their base versions. However, this evaluation may be influenced by stylistic alignment, favoring answers matching the judge LLM’s preferences. Furthermore, the judge LLM’s internal judgment mechanisms are unknown, making it unclear whether it prioritizes reasoning or stylistic coherence. Nevertheless, finetuned models achieve more correct answers than their base versions, and RAG models perform similarly, as confirmed by other evaluations. This raises questions about the sharpness of distinctions between correct and incorrect answers; more refined evaluation categories, similar to those in human evaluation, could provide clarity. Since this method correlates well with human evaluations for correct answers and scales efficiently, it is suitable for automated evaluation when sufficient contextual information is available.

Human evaluation is considered the gold standard but, unlike other quality metrics, cannot be scaled to large sample sizes. Nevertheless, even with 10-40 questions as in this work, it can reveal statistically significant performance differences. Due to the limited sample size, human evaluation typically does not comprehensively cover all relevant subdomains. Additionally, questions from a small group of experts introduce bias in question style. More semantically complex questions, as observed here, may advantage larger models with better general language understanding. Finally, we acknowledge that our evaluation scheme considers only factual correctness and does not involve direct pairwise preference evaluation, such as Elo scores [35]. Since the models already differ substantially in factual correctness, we consider pairwise comparisons a promising next step.

To summarize, this work demonstrates the importance of multi-faceted assessment approaches, as highlighted by [36]. No single evaluation method provides comprehensive coverage of model capabilities, with each approach offering distinct advantages and limitations. Our evaluation differs from previous studies that typically rely on one or two methods [3, 4, 17, 19, 20], revealing limitations that narrower evaluations may overlook. This concern was also emphasized by [36, 37]. Statistical significance analysis is crucial to our methodology, as all evaluation methods are influenced by

statistical variation. This leads to models achieving similar rankings despite different scores, making it difficult to establish clear performance hierarchies. Consequently, performance differences reported in other studies [4, 17, 19, 20] may be attributable to chance rather than meaningful distinctions, as cautioned in [37].

3.2 Methods for domain specialization

The results demonstrate that finetuning enhances factual knowledge in domain-specific tasks. In in-distribution tests, which consisted of multiple-choice evaluations and automated text similarity metrics, finetuned models often match or exceed larger state-of-the-art models like Llama 3.1 70B and Claude Sonnet 3.7, aligning with prior work [4, 11, 17–20, 36]. This confirms that domain-specific finetuning is effective in specialized areas like ECG without continual pretraining. However, human evaluation reveals limitations: Finetuned models underperform on questions syntactically different from training data, where smaller state-of-the-art or base models may excel. This likely results from insufficient complexity and diversity in the finetuning dataset [11]. Solutions include more complex training data or pretraining [17], though one study questions the effectiveness of the latter [37]. Additionally, synthetic data generation via LLM prompting lacks control over diversity and complexity, potentially limiting finetuning results. Human quality assurance [38] and more varied training data could address these issues. Overall, finetuning enables strong in-distribution performance but struggles with out-of-distribution queries due to limited data diversity, highlighting areas for refinement in synthetic data generation.

RAG-enhanced models perform similarly to finetuned models across most metrics. On in-distribution data, especially multiple-choice questions, they match finetuned models and can outperform larger models. However, they show notably decreased performance on BLEU, ROUGE, and BERTScore evaluations, supporting that lexical metrics are less meaningful in isolation. On out-of-distribution data in human evaluation, RAG consistently outperforms finetuned models and reaches Claude’s performance, aligning with [26]. As shown in previous work [16, 36], RAG offers a flexible option for knowledge enhancement. Notably, Llama 3.1 8B with RAG matches larger models while enabling local hosting for data control and independence from external dependencies—essential factors in medicine.

Claude Sonnet 3.7, representing large-scale proprietary models, is outperformed on multiple-choice, BLEU, ROUGE, and BERTScore evaluations but demonstrates strong performance in human evaluation. This shows that while imperfect on specific details, general-purpose models possess strong broad knowledge also in specialized domains such as ECG analysis. One study demonstrates that improvements in one domain can transfer to others without explicit training [39], which alongside likely extensive finetuning on PubMed data [37] may explain the strong performance of large-scale proprietary models.

Turning to a direct comparison, both RAG and finetuning increase ECG domain knowledge. Finetuning outperforms RAG on specialized data distributions, as demonstrated in automatic evaluations with syntactic and semantic constraints. However, performance remains similar in other in-distribution tests, indicating that further analysis is needed to determine exact differences. Human evaluations suggest base models

perform better on out-of-distribution queries. Production suitability depends on factors such as document accessibility, vector database availability, and intended use. RAG offers a clear advantage: easy updates to reflect evolving literature, whereas finetuning requires costly retraining.

To improve finetuned models, continual pretraining followed by supervised finetuning shows promise [17], though one study questions whether pretraining yields substantial improvements over state-of-the-art models [37]. A more varied training dataset covering diverse medical tasks (diagnoses, recommendations, complex relationships) and including human-generated data or complex chat instructions could significantly improve performance. Another option is finetuning with RLHF, where human feedback trains a reward model to guide the LLM toward defined preferences [40]. RAG implementation could also be improved through deeper analysis of embedding models, finetuning embeddings, or curating document selection. Combining RAG and finetuning warrants evaluation to leverage both approaches' advantages. Since finetuned models underperform on syntactically and semantically different questions, further investigation is needed to determine whether they can effectively use contextual information or require different finetuning strategies for RAG integration.

3.3 Clinical application scenarios

Our proposed models could enhance ECG interpretation accuracy and speed in clinical ECG practice, supporting immediate decision-making for tasks such as arrhythmia identification, ST-segment analysis, and QT assessment [21]. By providing instant, evidence-based guidance at the point of care, these tools could help reduce diagnostic errors and improve patient outcomes. Local deployment ensures data protection for patient health information while eliminating reliance on external searches, addressing both privacy concerns and workflow efficiency. Specific clinical applications demonstrate practical utility: (1) real-time emergency interpretation for distinguishing ST-elevation myocardial infarction (STEMI) mimics from true infarction, (2) continuous monitoring integration for intelligent alarm management and arrhythmic event prediction [41], and (3) rapid processing of large-scale ECG databases for novel biomarker identification [22]. Supporting evidence shows promise across cardiology tasks, with ChatGPT-4o achieving 75.9% accuracy on cardiology board questions [42]. These models could also integrate into advanced systems, including multimodal diagnostic frameworks or agentic systems for context-aware clinical reasoning, such as VLMs for ECG graph interpretation. Systems like ECG-Chat demonstrate this capability, combining ECG signal processing with textual analysis [29]. However, clinical deployment requires extensive validation including real-world testing, safety evaluations, workflow integration, and regulatory approval. At this stage, these models are better suited for augmenting clinical practice by helping clinicians find information rather than making autonomous decisions.

4 Methods

4.1 Finetuning (FT)

We generated question-answer pairs by prompting an LLM with literature context, similar to [38] (full prompt in Appendix A). Starting from PDF files, we converted them to markdown using MinerU [43], then cleaned the files by removing author information, tables of contents, references, and acknowledgements as in [44]. We split files by chapters (maximum 10 chapters and 50,000 tokens estimated with TikToken [45]) and prompted Llama 3.3 70B to generate Q&A pairs. We selected Llama 3.3 70B as an open-source model that consistently generated context-aligned questions without incorporating external knowledge, verified using AlignScore [46].

We finetuned Llama 3.1 70B and 8B models and compared them against their base versions and Claude Sonnet 3.7, representing state-of-the-art performance. We used the Hugging Face Trainer with QLoRA for memory-efficient training [47, 48]. The dataset was split 80/10/10 (train/validation/test) per file, with essential knowledge labeled by the human expert fully integrated into training. Prompts followed Llama 3 specifications with role tokens, bos, and eos tokens [25], with loss computed only on answers. We used AdamW optimizer (paged-32) [49], cosine learning-rate scheduler [50], batch size of eight, and cross-entropy loss [47]. LoRA parameters were applied to attention, feedforward, and output layers [51], resulting in 3.7% trainable parameters. Testing multiple configurations yielded optimal performance at $r = 256$, $\alpha = 128$ [50]. We monitored performance using domain-specific multiple-choice metrics (10% per file). Training loss decreased initially but validation loss increased after two epochs while multiple-choice accuracy stabilized, prompting us to stop training at two epochs. Higher α values (512) reduced 8B model accuracy. Experiments with Llama 3.1 70B on H100 GPUs showed similar patterns.

4.2 Retrieval-augmented generation (RAG)

RAG offers another approach for domain question-answering [16] and reduces LLM hallucination when knowledge is missing [3]. The RAG process involves several steps: First, documents converted to markdown using MinerU are split and embedded using a selected splitting strategy, vector database, and embedding model. During retrieval, the user’s message performs a similarity search on the vector database, selecting the top-k matching chunks as context. In augmentation, a prompt combines the user message and context, which is then passed to the LLM for response generation [15, 52].

We tested multiple configurations (detailed in Appendix C). The optimal configuration uses recursive splitting with 1024-token chunks and 100-token overlap [53], PubMedBERT embeddings [54], top-20 retrieval, and reranking with top-5, achieving first rank across all subsets. Reranking embeds document parts and queries simultaneously, enabling more precise similarity ranking [55]. We implemented this RAG configuration for Llama 3.1 8B Instruct and Llama 3.1 70B Instruct models.

4.3 Evaluation methodology

We assess the quality of the generated content by means of four qualitatively different evaluation methods, see Figure 5 for an overview, which we describe below.

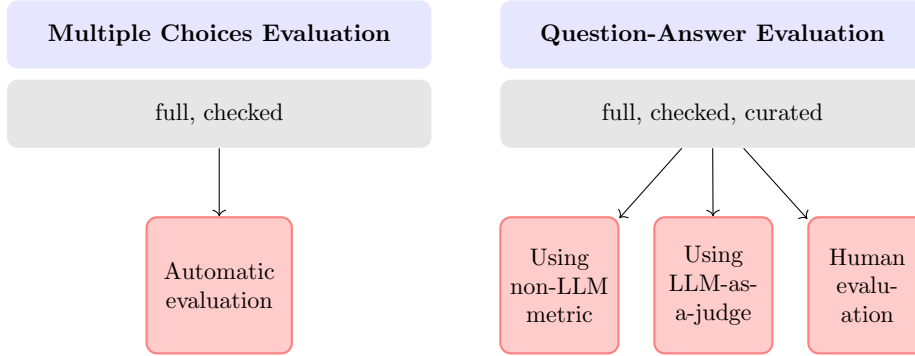


Fig. 5 Evaluation layers in the evaluation approach

Multiple-choice evaluation

Evaluating LLM knowledge requires a multi-faceted approach [36]. While multiple-choice tests can assess LLM understanding [17, 56], and general medical tests exist on Hugging Face [57], these are insufficient for evaluating specialized ECG knowledge. Therefore, we generated a domain-specific multiple-choice test from the provided medical documents using the data generation process for generating the finetuning data, adjusting the prompt to generate multiple-choice questions with answer options instead of question-answer pairs.

Text similarity metrics

Multiple-choice tests cannot represent real-world scenarios where LLMs answer open-ended questions. Consequently, a comprehensive evaluation should include both factual knowledge assessments and open-ended question evaluations. An evaluation based on automatic evaluation metrics supplements open-ended response evaluation despite uncertain quality assessment. Metrics like BLEU [31] and ROUGE [32], designed for summarization and translation, are suboptimal for question-answering but can assess keyword presence in responses. BERTScore [33] evaluates semantic meaning using contextual embeddings, accommodating synonymous expressions, though its quality depends on text formulation and cannot account for semantically valid alternative interpretations beyond reference answers. In our evaluation approach, we employ all of these metrics to provide a comprehensive evaluation of model responses.

LLM-as-a-judge

For open-ended questions, the answers are also assessed using LLM-as-a-judge [58]. We selected Deepseek R1 as the evaluator due to its strong reasoning capabilities

[59], making it effective for comparing responses and assessing accuracy. Deepseek R1 functions only as an evaluator and is not itself evaluated. The judge LLM focuses solely on correctness rather than ranking or scoring. Most importantly, we include the context that was used to originally generate the question in order to mitigate hallucinations [3]

Human evaluation

Human evaluation is carried out as a very valuable evaluation to confirm the other assessments. For this purpose, we collected the answers of finetuned models and state-of-the-art models to the questions provided by the medical expert and handed them over to the medical expert. The expert assesses whether the respective answers are correct, correct but incomplete, incorrect, or partially incorrect. We decided not to assign points to the answers themselves, as the allocation of points could be too variable and the criteria for evaluation could vary unconsciously. This process is carried out for two distinct subsets of questions provided by the human expert. Human evaluation is resources are limited and it is time-consuming [60].

Statistical significance testing

Statistical fluctuations may obscure true method rankings. We use empirical bootstrapping ($n = 1000$ iterations) to assess statistical significance. For each evaluation mode, we used respective scores as input; for human evaluation, we converted categorical answers to numerical scores (correct: 1; partially correct: 0.75; partially incorrect: 0.25; incorrect: 0). We performed pairwise comparisons by bootstrapping score differences between models. Performance differences are considered statistically significant if the 95% confidence interval excludes zero. This procedure generates ranked lists with ties, where ties indicate no statistically significant performance difference. This framework provides statistically robust, unified comparisons across all evaluation modes.

Evaluation datasets

For question-answer evaluation, we used a subset from the data generation process, split into train-validation-test sets (80/10/10), with test data serving as the evaluation dataset. While LLM-generated data lacks verified quality, it provides high quantity. These datasets (question-answer pairs and multiple-choice questions) are labeled as *full*. A medical expert corrected subsets of 534 multiple-choice and 537 question-answer samples, labeled as *checked*. For human evaluation, the expert additionally provided 47 questions, labeled as *curated*.

Declarations

Conflict of interest

The authors declare no competing interests.

Data availability

We release the human evaluation dataset along with model predictions and expert evaluations. The dataset used for pretraining and automatic evaluation cannot be released to align with §60d UrhG (implementing EU Directive 2019/790). The same holds for the model weights of the finetuned model.

Code availability

The complete source code for preprocessing, finetuning and retrieval augmented generation is available at <https://github.com/AI4HealthUOL/ecg-llm>.

Author contributions

Conceptualization: NS, WH; Methodology: NS, LA; Software: LA; Validation: LA, NS; Data Curation: WH; Visualization: LA; Writing - Original Draft: LA, NS; Writing - Review: LA, NS, WH; Supervision: NS

References

- [1] Kell, G., Roberts, A., Umansky, S., Qian, L., Ferrari, D., Soboczenski, F., Wallace, B.C., Patel, N., Marshall, I.J.: Question answering systems for health professionals at the point of care—a systematic review. *Journal of the american medical informatics association* **31**(4), 1009–1024 (2024)
- [2] Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A.: A comprehensive overview of large language models. *ACM Trans. Intell. Syst. Technol.* **16**(5) (2025) <https://doi.org/10.1145/3744746>
- [3] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., *et al.*: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* **43**(2), 1–55 (2025) [arXiv:2311.05232](https://arxiv.org/abs/2311.05232)
- [4] Li, R., Wang, X., Yu, H.: LlamaCare: An instruction fine-tuned large language model for clinical NLP. In: Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pp. 10632–10641. ELRA and ICCL, Torino, Italia (2024)
- [5] Jeong, D.P., Garg, S., Lipton, Z.C., Oberst, M.: Medical adaptation of large language and vision-language models: Are we making progress? In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12143–12170. Association for Computational Linguistics, Miami, Florida, USA (2024). <https://doi.org/10.18653/v1/2024.emnlp-main.677>

- [6] Wu, E., Wu, K., Zou, J.: Limitations of learning new and updated medical knowledge with commercial fine-tuning large language models. *NEJM AI* (2025) <https://doi.org/10.1056/aics2401155>
- [7] Ranjit, M., Srivastav, S., Ganu, T.: Radphi-3: Small language models for radiology. *arXiv preprint 2411.13604* (2024) [arXiv:2411.13604](https://arxiv.org/abs/2411.13604) [cs.CV]
- [8] Schilling-Wilhelmi, M., Ríos-García, M., Shabih, S., Gil, M.V., Miret, S., Koch, C.T., Márquez, J.A., Jablonka, K.M.: From text to insight: Large language models for materials science data extraction. *arXiv preprint 2407.16867* (2024) [arXiv:2407.16867](https://arxiv.org/abs/2407.16867) [cond-mat.mtrl-sci]
- [9] Toshniwal, S., Du, W., Moshkov, I., Kisacanin, B., Ayrapetyan, A., Gitman, I.: Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2025). <https://openreview.net/forum?id=mTCbq2QssD>
- [10] Yang, Z., Band, N., Li, S., Candès, E., Hashimoto, T.: Synthetic continued pre-training. In: *International Conference on Learning Representations (ICLR 2025)* (2025). <https://openreview.net/forum?id=07yvxWDSla>
- [11] Zhao, E., Awasthi, P., Haghtalab, N.: From style to facts: Mapping the boundaries of knowledge injection with finetuning. *arXiv preprint 2503.05919* (2025) [arXiv:2503.05919](https://arxiv.org/abs/2503.05919) [cs.CL]
- [12] Wu, E., Wu, K., Zou, J.: Finetunebench: How well do commercial fine-tuning apis infuse knowledge into llms? *arXiv preprint 2411.05059* (2024) [arXiv:2411.05059](https://arxiv.org/abs/2411.05059) [cs.CL]
- [13] Lin, J., Wang, Z., Qian, K., Wang, T., Srinivasan, A., Zeng, H., Jiao, R., Zhou, X., Gesi, J., Wang, D., et al.: Sft doesn't always hurt general capabilities: Revisiting domain-specific fine-tuning in llms. *arXiv preprint arXiv:2509.20758* (2025) [arXiv:2509.20758](https://arxiv.org/abs/2509.20758) [cs.CL]
- [14] Liu, J., Lin, J., Liu, Y.: How much can rag help the reasoning of llm? *arXiv preprint 2410.02338* (2024) [arXiv:2410.02338](https://arxiv.org/abs/2410.02338) [cs.CL]
- [15] Kamath, U., Keenan, K., Somers, G., Sorenson, S.: *Large Language Models: A Deep Dive: Bridging Theory and Practice*. Springer, Cham (2024). <https://doi.org/10.1007/978-3-031-65647-7>
- [16] Ong, C.S., Obey, N.T., Zheng, Y., Cohan, A., Schneider, E.B.: Surgeryllm: a retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. *npj Digital Medicine* **7**(1), 364 (2024)

- [17] Haghighi, T., Gholami, S., Sokol, J.T., Kishnani, E., Ahsaniyan, A., Rahmadian, H., Hedayati, F., Leng, T., Alam, M.N.: Eye-llama, an in-domain large language model for ophthalmology. *iScience* (2025) <https://doi.org/10.1016/j.isci.2025.112984>
- [18] Yang, Y., Tang, Y., Tam, K.Y.: Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint 2309.13064* (2023) [arXiv:2309.13064](https://arxiv.org/abs/2309.13064) [q-fin.GN]
- [19] Rosati, R., Antonini, F., Muralikrishna, N., Tonetto, F., Mancini, A.: Improving industrial question answering chatbots with domain-specific llms fine-tuning. In: 2024 20th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA), pp. 1–7 (2024). <https://doi.org/10.1109/MESA61532.2024.10704843>
- [20] Chao, C.-J., Banerjee, I., Arsanjani, R., Ayoub, C., Tseng, A., Delbrouck, J.-B., Kane, G.C., Lopez-Jimenez, F., Attia, Z., Oh, J.K., Erickson, B., Fei-Fei, L., Adeli, E., Langlotz, C.: Evaluating large language models in echocardiography reporting: opportunities and challenges. *European Heart Journal - Digital Health* **6**(3), 326–339 (2025) <https://doi.org/10.1093/ehjdh/ztae086>
- [21] Santos, J.F., Ladeiras-Lopes, R., Leite, F., Dores, H.: Applications of large language models in cardiovascular disease: A systematic review. *European Heart Journal - Digital Health* **6**(4), 540–550 (2025) <https://doi.org/10.1093/ehjdh/ztaf028>
- [22] Losch, N., Plagwitz, L., Büscher, A., Varghese, J.: Fine-tuning llms on small medical datasets: Text classification and normalization effectiveness on cardiology reports and discharge records. *arXiv preprint 2503.21349* (2025) [arXiv:2503.21349](https://arxiv.org/abs/2503.21349) [cs.CL]
- [23] Jung, H., Kim, Y., Choi, H., Seo, H., Kim, M., Han, J., Kee, G., Park, S., Ko, S., Kim, B., Kim, S., Jun, T.J., Kim, Y.-H.: Enhancing clinical efficiency through llm: Discharge note generation for cardiac patients. *arXiv preprint 2404.05144* (2024) [arXiv:2404.05144](https://arxiv.org/abs/2404.05144) [cs.CL]
- [24] Lee, P.C., Sharma, S.K., Motaganahalli, S., Huang, A.: Evaluating the clinical decision-making ability of large language models using mksap-19 cardiology questions. *JACC: Advances* **2**(9), 100658 (2023) <https://doi.org/10.1016/j.jacadv.2023.100658>
- [25] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C.C., Nikolaidis, C., Allonsius, D., Song, D., Pintz,

D., Livshits, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I., Misra, I., Evtimov, I., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K.V., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Rantala-Yeary, L., Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Tan, X.E., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Grattafiori, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Vaughan, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Franco, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B.D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Wyatt, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Ozgenel, F., Caggioni, F., Guzmán, F., Kanayet, F., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Thattai, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Damla, I., Molybog, I., Tufanov, I., Veliche, I.-E.,

- Gat, I., Weissman, J., Geboski, J., Kohli, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K.H., Saxena, K., Prasad, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Huang, K., Chawla, K., Lakhotia, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Tsimpoukelli, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Laptev, N.P., Dong, N., Zhang, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Li, R., Hogan, R., Battey, R., Wang, R., Maheswari, R., Howes, R., Rinott, R., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Kohler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Albiero, V., Ionescu, V., Poenaru, V., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wang, X., Wu, X., Wang, X., Xia, X., Wu, X., Gao, X., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Hao, Y., Qian, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z.: The Llama 3 Herd of Models (2024)
- [26] Radke, R.M., Diller, G.-P., Reddy, R.G., Shivaram, P., Danford, D.A., Kutty, S.: A multi-query, multimodal, receiver-augmented solution to extract contemporary cardiology guideline information using large language models. *European Heart Journal - Digital Health*, 111 (2025) <https://doi.org/10.1093/ehjdh/ztaf111> <https://academic.oup.com/ehjdh/advance-article-pdf/doi/10.1093/ehjdh/ztaf111/64356052/ztaf111.pdf>
- [27] Zhou, Y., Zhang, P., Song, M., Zheng, A., Lu, Y., Liu, Z., Chen, Y., Xi, Z.: Zodiac: A cardiologist-level llm framework for multi-agent diagnostics. *arXiv preprint 2410.02026* (2024) [arXiv:2410.02026](https://arxiv.org/abs/2410.02026) [cs.AI]
- [28] Zhang, Y., Bunting, K.V., Champai, A., Wang, X., Lu, W., Thorley, A., Hothi, S.S., Qiu, Z., Kotecha, D., Duan, J.: Cardiac-agents: A multimodal framework with hierarchical adaptation for cardiac care support (2025) [arXiv:2508.13256](https://arxiv.org/abs/2508.13256) [cs.AI]

- [29] Zhao, Y., Kang, J., Zhang, T., Han, P., Chen, T.: Ecg-chat: A large ecg-language model for cardiac disease diagnosis. arXiv preprint 2408.08849 (2025) [arXiv:2408.08849](https://arxiv.org/abs/2408.08849) [eess.SP]
- [30] Pham, H.M., Tang, J., Saeed, A., Ma, D.: Q-heart: Ecg question answering via knowledge-informed multimodal llms. arXiv preprint 2505.06296 (2025) [arXiv:2505.06296](https://arxiv.org/abs/2505.06296) [eess.SP]
- [31] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, USA (2002). <https://doi.org/10.3115/1073083.1073135>
- [32] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004)
- [33] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020)
- [34] Mohammed, L.A., Aljaberi, M.A., Anmary, A.S., Abdulkhaleq, M.: Analysing english for science and technology reading texts using flesch reading ease online formula: The preparation for academic reading. In: Al-Sharafi, M.A., Al-Emran, M., Al-Kabi, M.N., Shaalan, K. (eds.) Proceedings of the 2nd International Conference on Emerging Technologies and Intelligent Systems, pp. 546–561. Springer, Cham (2023)
- [35] Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J.E., *et al.*: Chatbot arena: An open platform for evaluating llms by human preference. In: Forty-first International Conference on Machine Learning (2024)
- [36] Teo, Z.L., Thirunavukarasu, A.J., Elangovan, K., Cheng, H., Moova, P., Soetikno, B., Nielsen, C., Pollreisz, A., Ting, D.S.J., Morris, R.J.T., Shah, N.H., Langlotz, C.P., Ting, D.S.W.: Generative artificial intelligence in medicine. Nature medicine (2025) <https://doi.org/10.1038/s41591-025-03983-2>
- [37] Jeong, D.P., Garg, S., Lipton, Z.C., Oberst, M.: Medical adaptation of large language and vision-language models: Are we making progress? In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 12143–12170. Association for Computational Linguistics, Miami, Florida, USA (2024). <https://doi.org/10.18653/v1/2024.emnlp-main.677>
- [38] Zheng, Q., Abdullah, S., Rawal, S., Zakka, C., Ostmeier, S., Purk, M., Reis, E.,

- Topol, E.J., Leskovec, J., Moor, M.: Miriad: Augmenting llms with millions of medical query-response pairs. arXiv preprint 2506.06091 (2025) [arXiv:2506.06091](#) [cs.CL]
- [39] Li, Y., Pan, Z., Lin, H., Sun, M., He, C., Wu, L.: Can one domain help others? a data-centric study on multi-domain reasoning via reinforcement learning. arXiv preprint 2507.17512 (2025) [arXiv:2507.17512](#) [cs.AI]
- [40] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA (2022)
- [41] Khera, R., Oikonomou, E.K., Nadkarni, G.N., *et al.*: Transforming cardiovascular care with artificial intelligence: From discovery to practice. *Journal of the American College of Cardiology* **84**(2), 97–114 (2024) <https://doi.org/10.1016/j.jacc.2024.05.003>
- [42] Malik, A., Madias, C., Wessler, B.S.: Performance of chat generative pre-trained transformer-4o in the adult clinical cardiology self-assessment program. *European Heart Journal - Digital Health* **6**(1), 155–158 (2025) <https://doi.org/10.1093/ehjdh/ztae077>
- [43] Wang, B., Xu, C., Zhao, X., Ouyang, L., Wu, F., Zhao, Z., Xu, R., Liu, K., Qu, Y., Shang, F., Zhang, B., Wei, L., Sui, Z., Li, W., Shi, B., Qiao, Y., Lin, D., He, C.: Mineru: An open-source solution for precise document content extraction. arXiv preprint 2409.18839 (2024) [arXiv:2409.18839](#) [cs.CV]
- [44] Schilling-Wilhelmi, M., Ríos-García, M., Shabih, S., Gil, M.V., Miret, S., Koch, C.T., Márquez, J.A., Jablonka, K.M.: From Text to Insight: Large Language Models for Materials Science Data Extraction (2024). <http://arxiv.org/pdf/2407.16867>
- [45] OpenAI: tiktoken: Fast Byte Pair Encoding Tokenizer. <https://github.com/openai/tiktoken>. Accessed: 2025-02-08 (2023)
- [46] Zha, Y., Yang, Y., Li, R., Hu, Z.: Alignscore: Evaluating factual consistency with a unified alignment function. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 11328–11348. Association for Computational Linguistics, Toronto, Canada (2023)
- [47] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q.,

- Rush, A.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, Online (2020). <https://doi.org/10.18653/v1/2020.emnlp-demos.6> . <https://aclanthology.org/2020.emnlp-demos.6/>
- [48] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: Qlora: Efficient fine-tuning of quantized llms. *Advances in neural information processing systems* **36**, 10088–10115 (2023) [arXiv:2305.14314](https://arxiv.org/abs/2305.14314) [cs.LG]
 - [49] Marie, B.: Fine-tuning LLMs with 32-bit, 8-bit, and Paged AdamW Optimizers. Accessed: 2025-06-14 (2024). <https://kaiichup.substack.com/p/fine-tuning-llms-with-32-bit-8-bit>
 - [50] Raschka, S.: Practical Tips for Finetuning LLMs Using LoRA. Accessed: 2025-06-01 (2024). <https://magazine.sebastianraschka.com/p/practical-tips-for-finetuning-llms>
 - [51] Hu, E.J., shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
 - [52] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 9459–9474 (2020) [arXiv:2005.11401](https://arxiv.org/abs/2005.11401) [cs.CL]
 - [53] Theja, R.: Evaluating the ideal chunk size for a rag system using llamaindex. *LlamaIndex Blog* (2023). Accessed: 2025-05-03
 - [54] NeuML: PubMedBERT Base Embeddings. Accessed: 2025-05-26 (2023). <https://huggingface.co/NeuML/pubmedbert-base-embeddings>
 - [55] Ampazis, N.: Improving rag quality for large language models with topic-enhanced reranking. In: IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 74–87 (2024). Springer
 - [56] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. In: International Conference on Learning Representations (2021)
 - [57] Ura, A., Minervini, P., Fourrier, C.: The Open Medical-LLM Leaderboard: Benchmarking Large Language Models in Healthcare. Accessed: 2025-04-29 (2024). <https://huggingface.co/blog/leaderboard-medicalllm>
 - [58] Kim, S., Suk, J., Longpre, S., Lin, B.Y., Shin, J., Welleck, S., Neubig, G., Lee, M., Lee, K., Seo, M.: Prometheus 2: An open source language model specialized

- in evaluating other language models. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353. Association for Computational Linguistics, Miami, Florida, USA (2024). <https://doi.org/10.18653/v1/2024.emnlp-main.248>
- [59] DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z.F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J.L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R.J., Jin, R.L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S.S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W.L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X.Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y.K., Wang, Y.Q., Wei, Y.X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y.X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z.Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., Zhang, Z.: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (2025)
- [60] Li, T., Chiang, W.-L., Frick, E., Dunlap, L., Wu, T., Zhu, B., Gonzalez, J.E., Stoica, I.: From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In: *Forty-second International Conference on Machine Learning* (2025)
- [61] Pal, A., Minervini, P., Motzfeldt, A.G., Gema, A.P., Alex, B.: The Open Medical-LLM Leaderboard: Benchmarking Large Language Models in Healthcare. https://huggingface.co/spaces/openlifescienceai/open_medical_llm_leaderboard. Hugging Face (2024)
- [62] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* **3**(1), 1–23 (2021) [arXiv:2007.15779](https://arxiv.org/abs/2007.15779)

Appendix A Prompt data generation

Listing 1 shows the prompt for the generation of Q&A-pairs

```
1 system_message = f"""You are a Teacher/ Professor in the medical
   field. Your task is to setup a examination with free text
   answers. Using the provided context, formulate questions with
   different difficulties that capture the medical content from
   the context. Please also give the answers, so that the test
   could be corrected afterwards.
2 If you cannot generate any question to medical content, please
   skip it.
3 You MUST obey the following criteria:
4 - Restrict the question to the context information provided.
5 - vary between different question words (e.g. what, why, which,
   how, etc.)
6 - Ensure every question is fully self-contained and answerable
   without requiring additional context or prior
   questions/answers.
7 - Do NOT ask for figures, algortihms, tables, names of the
   present study or similar.
8 - Do NOT put phrases like "given provided context" or "in this
   work" or "in this case" or "what is algorithm a" or some
   questions regarding the study
9 - Replace these terms with specific details
10 - ONLY ask for medical details. DO NOT ask about the study,
   author or index, IGNORE them
11
12 BAD questions:
13 - What are the symptoms of the disease described
14 - How many patients were included in the study
15
16 GOOD questions:
17 - What are the symptoms of Corona
18 - Why should the patient drink so much water when having a fever
19
20 Sometimes the context may contain overhead such as titles,
   authors, study information or similar. Please only use the
   content that has a medical context.
21
22 Output: ONLY JSON."""
23
24 user_message = f"""Here is the context for generating medical
   questions:
25
26 {context}
27 {self.parser.get_format_instructions()}
28 Please answer in English and do not use any German words or
   phrases. Translate technical terms into English if necessary.
29 """
```


Listing 1 Prompt to generate question and answer pairs

Appendix B Prompt generating multiple choices

Listing 2 shows the prompt for generating multiple choices.

```
1 system_message = f"""You are a Teacher/Professor in the medical
   field.
2 Your task is to create multiple-choice questions with different
   difficulties based on the medical context provided.
3
4 For each question:
5 - Generate one question about important medical content
6 - Create exactly 4 answer options
7 - One option must be correct and from the context
8 - Three options must be plausible but incorrect (make these up)
9 - All options should have similar length and style
10 - Focus only on medical content
11 If you cannot generate any question to medical content, please
   skip it.
12 You MUST obey the following criteria:
13 - Restrict the question to the context information provided.
14 - vary between different question words (e.g. what, why, which,
   how, etc.)
15 - Ensure every question is fully self-contained and answerable
   without requiring additional context or prior
   questions/answers.
16 - Do NOT ask for figures, algorithms, tables, names of the
   present study or similar.
17 - Do NOT put phrases like "given provided context" or "in this
   work" or "in this case" or "what is algorithm a" or some
   questions regarding the study
18 - Replace these terms with specific details
19 - ONLY ask for medical details. DO NOT ask about the study,
   author or index, IGNORE them
20
21 BAD questions:
22 - What are the symptoms of the disease described
23 - How many patients were included in the study
24
25 GOOD questions:
26 - What are the symptoms of Corona
27 - Why should the patient drink so much water when having a fever
28
29 Sometimes the context may contain overhead such as titles,
   authors, study information or similar. Please only use the
   content that has a medical context.
30
31 Output: ONLY JSON."""
```

```

32
33 user_message = f"""Generate multiple-choice questions from this
    context:
34 {context}
35 {parser.get_format_instructions()}
36 Please answer in English and do not use any German words or
    phrases. Translate technical terms into English if necessary.
37
38 """

```

Listing 2 Prompt to generate multiple-choices

Appendix C Results of different RAG-configurations

We considered various RAG configurations before selecting the final one. Therefore, we evaluated the multiple-choice results on the different subsets with different configurations. For splitting the documents, we consider the Markdown splitting approach as implemented in Langchain as the documents are converted as markdown files, and a recursive chunking strategy with a size of 1024 with 100 overlap suggested by [53]. For the embedding process, we selected Chroma as the vector database and compared two embedding models: multilingual-e5-large-instruct as one of the leading embedding models in the Hugging Face leaderboard for embedding models [61], and PubMedBERT [62] for its specialization in biomedical text. Table C1 shows the results of the different RAG configurations across splitting strategies, retrieval settings, and embedding models, focusing on factual accuracy in multiple-choice evaluations. Overall, recursive splitting with 1,024-token chunks, larger top-k values, and reranking achieves the best performance, while both embedding models performed similarly, with PubMedBERT showing slightly better performance in the checked subset.

Appendix D LLM-as-a-judge evaluation

To verify the LLM-as-a-judge evaluation, we gave a smaller, random subset of correct and incorrect labeled answers from one of the finetuned models (LLama 3.1 8B + FT, $r = 64, \alpha = 16$) to the human expert. An overview of the results is shown in Table D2. The correct answers are also labeled as correct. From the wrong answers, three answers are labeled as correct, one as partially correct.

Splitting	Top-k	Embedding	Full	Checked	English
Recursive (1,024 + 100)	20 + Reranking top 5	PubMedBERT	88.5% (1)	85.0% (1)	88.3% (1)
Recursive (1,024 + 100)	20 + Reranking top 5	Multilingual	88.8% (1)	83.9% (2)	88.5% (1)
Recursive (1,024 + 100)	10 + Reranking top 5	PubMedBERT	87.3% (3)	83.9% (3)	86.9% (5)
Recursive (1,024 + 100)	10 + Reranking top 5	Multilingual	87.6% (3)	82.8% (4)	87.2% (3)
Recursive (1,024 + 100)	10	Multilingual	87.5% (3)	82.7% (4)	87.2% (3)
Recursive (1,024 + 100)	5	PubMedBERT	86.7% (6)	82.0% (6)	86.3% (6)
Recursive (1,024 + 100)	5	Multilingual	86.7% (7)	81.1% (8)	86.5% (7)
Recursive (1,024 + 100)	10	PubMedBERT	86.5% (8)	82.2% (6)	86.1% (8)
Recursive (1,024 + 100)	3	Multilingual	85.9% (9)	80.5% (9)	85.7% (9)
Markdownheader	2	PubMedBERT	85.2% (10)	79.4% (10)	84.7% (10)
Markdownheader	2	Multilingual	84.4% (11)	76.2% (12)	84.0% (11)
Recursive (1,024 + 100)	3	PubMedBERT	83.9% (12)	78.7% (11)	83.7% (11)

Table C1 Combined and ranked comparison of RAG configurations showing multiple-choice accuracy with corresponding rank across subsets.

Table D2 Validation of LLM-as-a-judge evaluation through human expert review. Values represent counts of responses. Agreement between automated and human evaluation was 80% (16/20 cases). Cases labeled as incorrect by the LLM judge showed heterogeneous human assessments, highlighting the complexity of response quality evaluation.

LLM-as-a-Judge Label	Total Count	Human: Incorrect	Human: Correct	Human: Partially Correct
Correct	13	0	13	0
Incorrect	7	3	3	1

Appendix E Results of other tested configurations

We additionally finetuned models using various LoRA configurations. Specifically, we tested configurations with $\alpha \in 128, 256, 512$ and a the corresponding rank of $r = 256$. Furthermore, we evaluated a setup with $r = 64$ and $\alpha = 16$, as suggested by [4]. One experiment was also conducted using a training dataset from which all duplicated questions were removed. The results of these experiments are presented in the following Tables E3, E4, E5, E6 and Figures E1, E2, E3. For the 8B model, the configuration with $\alpha = 512$ failed to run successfully. Overall, the configuration with $r = 256$ and $\alpha = 128$ achieved the best performance. As initial experiments without duplicated questions resulted in worse performance compared to those including some duplicates, further testing under that condition was discontinued. The overall ranking is shown in Section F

Multiple choice evaluation

Table 1 presents the multiple-choice results for all finetuned configurations.

Model	Special	Full	Checked
LLama 3.1 8B Instruct	73.3%	81.6%	73%
LLama 3.1 70B Instruct	82.4%	88.2%	81.5%
Claude 3.7 Sonnet Latest	82.3%	88%	81.7%
LLama 3.1 8B + FT ($r=64, \alpha=16$)	83.7%	87.9%	83%
LLama 3.1 8B + FT ($r=64, \alpha=16$ no duplicates)	83.9%	87.6%	81.5%
LLama 3.1 8B + FT ($r=256, \alpha=128$)	84.9%	87.5%	84.3%
LLama 3.1 8B + FT ($r=256, \alpha=256$)	84.2%	85%	84.1%
LLama 3.1 8B + FT ($r=256, \alpha=512$)	79.3%	79.2%	80%
LLama 3.1 70B + FT ($r=256, \alpha=128$)	90.2%	92%	88.2%
LLama 3.1 70B + FT ($r=256, \alpha=256$)	87.4%	88.4%	84.8%
LLama 3.1 70B + FT ($r=256, \alpha=512$)	89.2%	92.1%	89.7%
LLama 3.1 8B + RAG	86.1%	88.5%	85%
Llama 3.1 70B + RAG	87.9%	90.2%	86.5%

Table E3 Performance of different models across different multiple-choice sets

BLEU, ROUGE, BERTScore evaluation

Table E4 presents the BLEU and ROUGE-1 values, table E5 the ROUGE-2 and ROUGE-L and Table E6 the BERTScores for all finetuned configurations.

LLM-as-a-judge evaluation

Figure E1 shows the results of LLM-as-a-judge for all finetuned configurations.

Model	BLEU	R1 F1	R1 Precision	R1 Recall
LLama 3.1 8B + FT ($r = 64, \alpha = 16$)	0.1103	0.4031	0.4035	0.4549
LLama 3.1 8B + FT ($r = 64, \alpha = 16$ no duplicates)	0.1083	0.4002	0.3984	0.4545
LLama 3.1 8B + FT ($r = 256, \alpha = 128$)	0.1195	0.4063	0.4039	0.4597
LLama 3.1 8B + FT ($r = 256, \alpha = 256$)	0.1196	0.4030	0.3996	0.4553
LLama 3.1 8B + FT ($r = 256, \alpha = 512$)	0.1149	0.3910	0.3898	0.4396
LLama 3.1 70B + FT ($r = 256, \alpha = 128$)	0.1289	0.4270	0.4292	0.4749
LLama 3.1 70B + FT ($r = 256, \alpha = 256$)	0.1355	0.4162	0.4168	0.4617
LLama 3.1 70B + FT ($r = 256, \alpha = 512$)	0.1281	0.4266	0.4262	0.4781
LLama 3.1 8B Instruct	0.0788	0.3537	0.2903	0.5252
Claude 3.7 Sonnet Latest	0.0694	0.3661	0.3247	0.4897
LLama 3.1 70B Instruct	0.0962	0.3852	0.3307	0.5347
LLama 3.1 8B Instruct + RAG	0.0763	0.3481	0.2856	0.5162
LLama 3.1 70B Instruct + RAG	0.0339	0.2414	0.1668	0.5488

Table E4 BLEU and ROUGE-1 scores for different models on test dataset

Model	R2 F1	R2 P	R2 R	RL F1	RL P	RL R
LLama 3.1 8B + FT ($r = 64, \alpha = 16$)	0.2197	0.2202	0.2483	0.3518	0.3523	0.3979
LLama 3.1 8B + FT ($r = 64, \alpha = 16$) no duplicates	0.2164	0.2149	0.2465	0.3482	0.3466	0.3967
LLama 3.1 8B + FT ($r = 256, \alpha = 128$)	0.2279	0.2263	0.2582	0.3568	0.3546	0.4046
LLama 3.1 8B + FT ($r = 256, \alpha = 256$)	0.2262	0.2239	0.2559	0.3538	0.3506	0.4004
LLama 3.1 8B + FT ($r = 256, \alpha = 512$)	0.2167	0.2153	0.2436	0.3443	0.3429	0.3878
LLama 3.1 70B + FT ($r = 256, \alpha = 128$)	0.2449	0.2455	0.2725	0.3764	0.3782	0.4192
LLama 3.1 70B + FT ($r = 256, \alpha = 256$)	0.2426	0.2424	0.2696	0.3694	0.3699	0.4103
LLama 3.1 70B + FT ($r = 256, \alpha = 512$)	0.2445	0.2437	0.2739	0.3761	0.3755	0.4226
LLama 3.1 8B Instruct	0.1730	0.1429	0.2586	0.2906	0.2382	0.4367
Claude 3.7 Sonnet Latest	0.1685	0.1502	0.2297	0.2992	0.2650	0.4054
LLama 3.1 70B Instruct	0.1981	0.1715	0.2766	0.3208	0.2750	0.4503
LLama 3.1 8B Instruct + RAG	0.1689	0.1397	0.2514	0.2831	0.2320	0.4250
LLama 3.1 70B Instruct + RAG	0.1064	0.0734	0.2512	0.1871	0.1288	0.4382

Table E5 ROUGE-2 and ROUGE-L scores for different models on test dataset

Human evaluation

Figure E2 shows the results of all finetuned models on the 10 questions asking for facts. Figure E3 shows the results of all finetuned models on the 40 more complex questions, evaluated by an human expert.

Model	Precision	Recall	F1
LLama 3.1 8B + FT ($r = 64, \alpha = 16$)	0.3644	0.3693	0.3665
LLama 3.1 8B + FT ($r = 64, \alpha = 16$) no duplicates	0.3610	0.3670	0.3636
LLama 3.1 8B + FT ($r = 256, \alpha = 128$)	0.3660	0.3748	0.3700
LLama 3.1 8B + FT ($r = 256, \alpha = 256$)	0.3599	0.3710	0.3651
LLama 3.1 8B + FT ($r = 256, \alpha = 512$)	0.3537	0.3573	0.3552
LLama 3.1 70B + FT ($r = 256, \alpha = 128$)	0.3884	0.3932	0.3904
LLama 3.1 70B + FT ($r = 256, \alpha = 256$)	0.3732	0.3826	0.3776
LLama 3.1 70B + FT ($r = 256, \alpha = 512$)	0.3846	0.3961	0.3900
LLama 3.1 8B Instruct	0.2194	0.3699	0.2936
Claude 3.7 Sonnet Latest	0.2703	0.4112	0.3395
LLama 3.1 70B Instruct	0.2795	0.4180	0.3476
LLama 3.1 8B Instruct + RAG	0.2157	0.3584	0.2861
LLama 3.1 70B Instruct + RAG	0.2811	0.1418	0.0095

Table E6 BERTScore of the models over test dataset

Appendix F Overall ranking

Table F7 shows all models with their corresponding ranks in the different evaluations.

Model	Full mult. choi.	Special multiple choices	Checked multiple choices	BLEU	R-1	R-2	R-L	BERT-Score	LLM-as-a-judge	Human 10 questions	Human 40 questions
LLama 3.1 8B	12	12	12	10	11	10	11	11	11	7	5
LLama 3.1 70B	6	7	9	9	9	9	9	9	3	5	1
Claude 3.7 Sonnet	6	7	9	12	10	11	10	10	1	1	1
LLama 3.1 8B + FT (r=64, $\alpha=16$)	6	7	5	7	6	6	5	5	5	7	8
LLama 3.1 8B + FT (r=64, $\alpha=16$ no duplicates)	9	7	9	7	5	7	7	6	6	7	
LLama 3.1 8B + FT (r = 256, $\alpha = 128$)	9	5	5	4	4	4	4	4	5	7	5
LLama 3.1 8B + FT (r = 256, $\alpha = 256$)	11	7	5	4	6	4	6	6	5	13	10
LLama 3.1 8B + FT (r = 256, $\alpha = 512$)	12	12	12	6	8	8	7	8	11	12	10
LLama 3.1 70B + FT (r = 256, $\alpha = 128$)	1	1	1	1	1	1	1	1	2	4	5
LLama 3.1 70B + FT (r = 256, $\alpha = 256$)	5	4	3	1	3	1	3	3	5	7	10
LLama 3.1 70B + FT (r = 256, $\alpha = 512$)	2	1	1	3	1	1	1	1	3	5	8
LLama 3.1 8B + RAG	4	5	5	11	12	11	12	12	5	1	1
LLama 3.1 70B + RAG	2	3	3	13	13	13	13	13	3	1	1

Table F7 Ranks of different models across the different evaluation methods

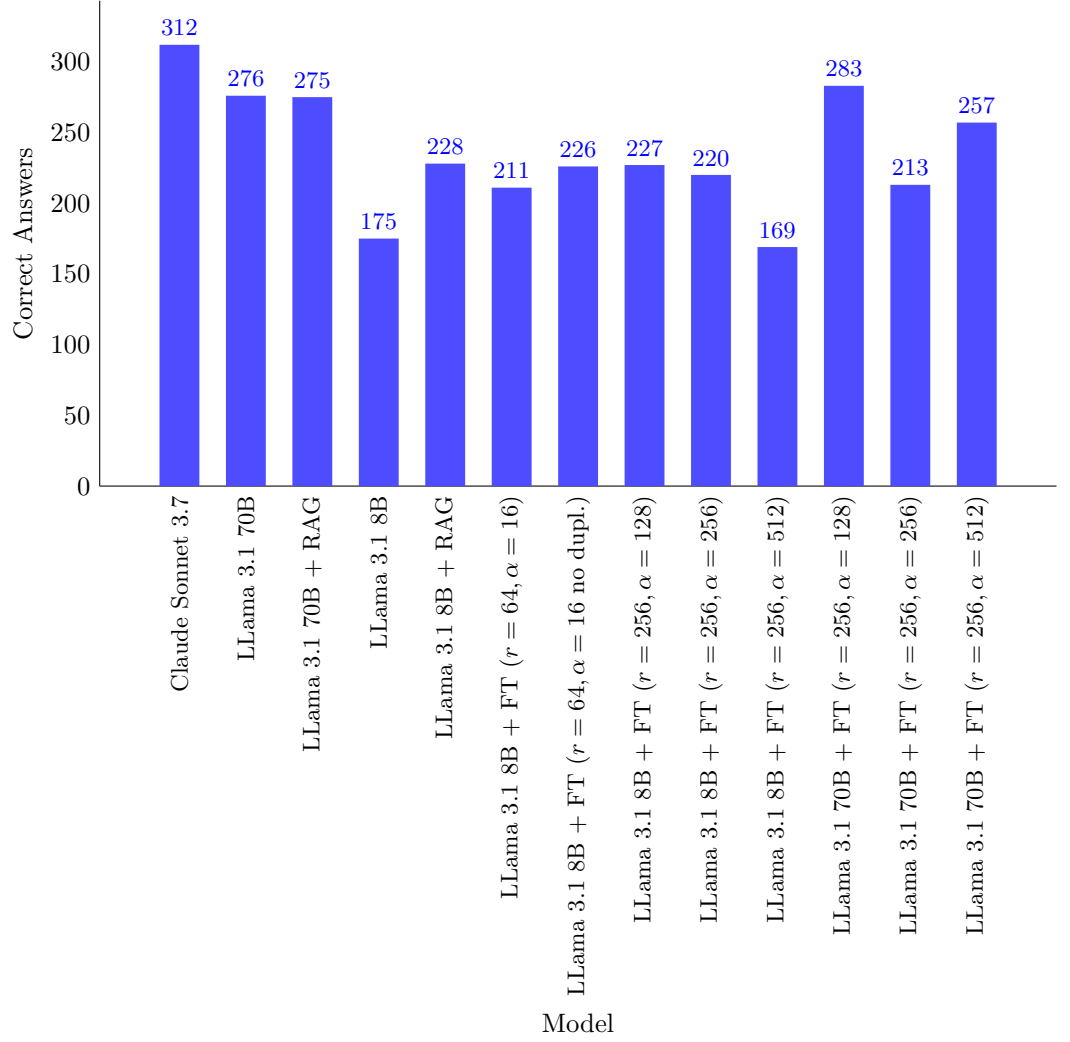


Fig. E1 Correct answers resulting from 416 questions evaluated with LLM-as-a-judge

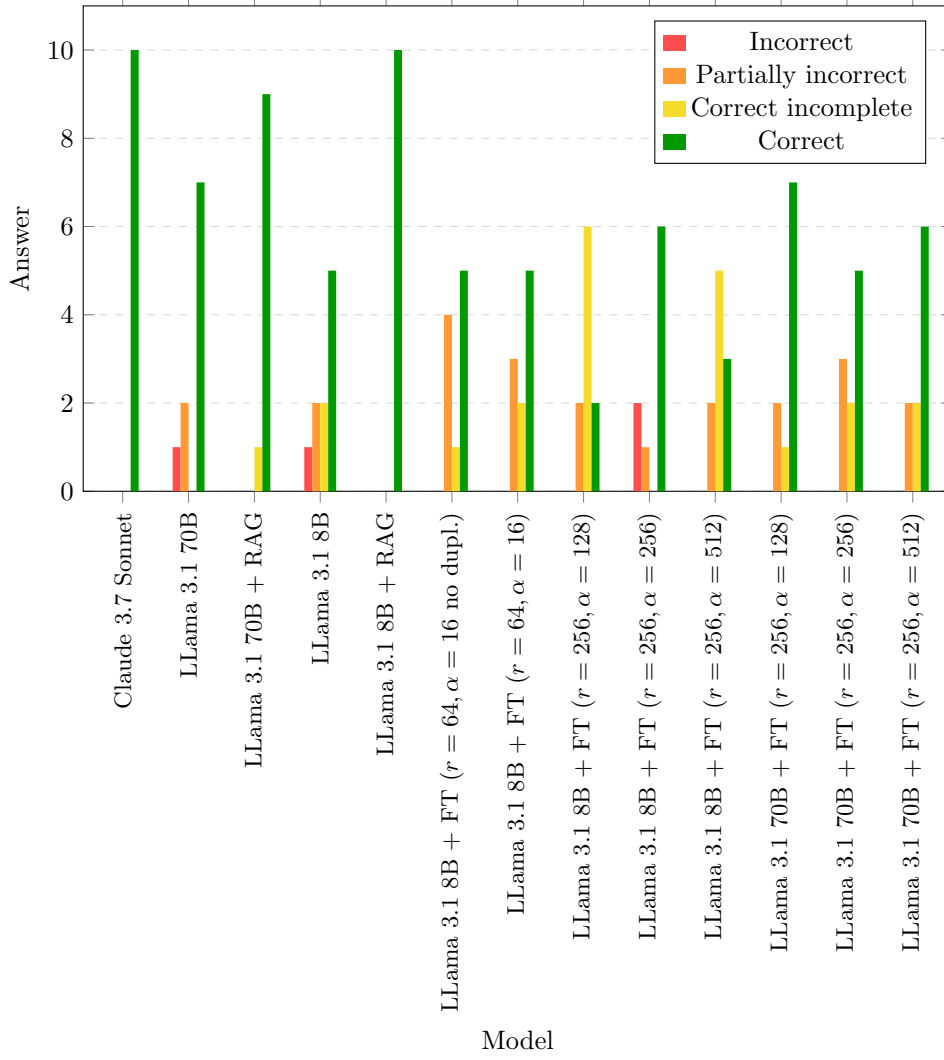


Fig. E2 Human evaluation on the ten factual questions

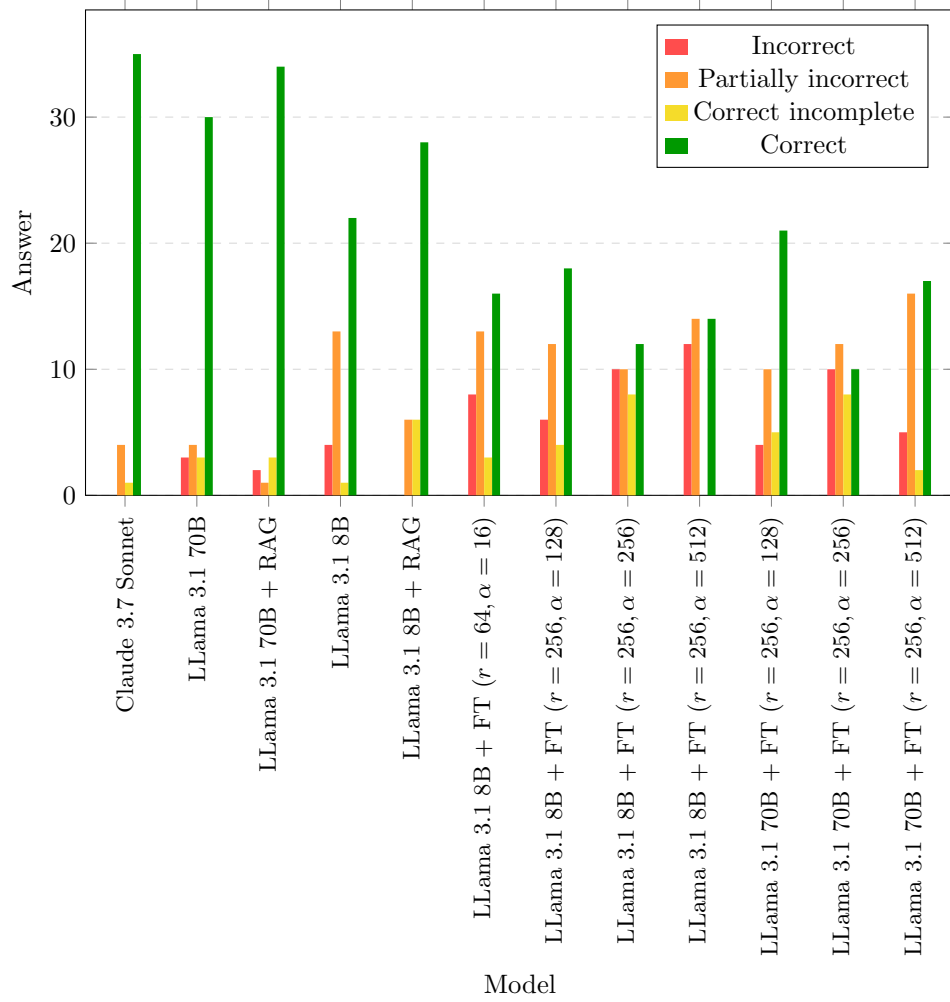


Fig. E3 Human evaluation on the 40 semantically complex questions