Group Relative Policy Optimization for Spanish Clinical Case Report Summarization

Notebook for the BioASQ Lab at CLEF 2025

Georgi Grazhdanski

Abstract

This paper evaluates a reinforcement-learning-based approach to summarizing clinical case reports in Spanish for the MultiClinSum shared task, part of the BioASQ workshop. We train a large language model using Group Relative Policy Optimization (GRPO) on a set of 500 full-text and summary pairs, and analyze how it compares against standard fine-tuning and domain pre-training methods. Our best system achieves 0.2899 ROUGE-L F1 and 0.7578 BERTScore F1 on the official challenge test set.

Keywords

Clinical text summarization, Spanish clinical text, Large language models, Reinforcement learning

1. Introduction

Clinical case reports play an important role in advancing medical knowledge by providing clinicians with the latest findings on adverse effects of treatments, rare diseases, unusual representation of known conditions, and many more. Some reports, however, can be rather lengthy, making it difficult for medical professionals to keep up with a large number of cases. Consequently, developing an automated system for faithfully summarizing such texts could be beneficial to the medical community, with potential applications in medical literature review, clinical decision support, clinical trial screening, and more.[1]

To address this, we explore several LLM-based approaches to automatic summarization of clinical case reports in Spanish, as part of the Multilingual Clinical Documents Summarization (MultiClinSum) shared task[1] in BioASQ 2025[2]. The MultiClinSum task organizers provide corpora of full-text clinical case reports and their corresponding human-generated summaries in English, French, Spanish, and Portuguese. Given the full text of a clinical case report, participating systems are required to generate a summary that effectively captures the key information from the source document.

Our experiments on the Spanish dataset show that training a Llama 3.1 8B Instruct[3] model using Group Relative Policy Optimization (GRPO)[4] can improve clinical summarization metrics, outperforming the standard supervised fine-tuning approach, as well as GPT-4.1 and a biomedical Llama 3 8B Instruct models in terms of ROUGE-L F1[5] and BERTScore F1[6] scores, in a 0-shot setting.

The code for the experiments can be found on GitHub[1].

2. Related Work

Large language models (LLMs) have shown promising results in medical NLP tasks, including summarization of various types of clinical documents. The authors in [7] explore the capabilities of LLMs to summarize radiology reports, patient questions, progress notes, and doctor-patient dialogue. They find that human experts often prefer summaries generated by the best-adapted LLMs to human-generated summaries, in terms of completeness and correctness.

Adapting LLMs for the medical domain has been shown to, in general, improve performance on downstream tasks [8], including clinical text summarization[7]. The authors in [9] fine-tune a Llama

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

georgi.grazhdanski@gmail.com (G. Grazhdanski)

0009-0008-7084-6788 (G. Grazhdanski)

[1] https://github.com/ggrazh/multiclinsum

3 8B model on a medical dialogue corpus (ACI-BENCH)[10] for the task of automatically generating medical reports from medical dialogues. They observe a significant improvement in terms of ROUGE1 and BERTScore compared to the unmodified Llama 3 8B Instruct model.

General large language models such as GPT-4 also show great results in clinical tasks[11]. The authors in [11] evaluate GPT-4 on a radiology findings summarization task, showing that GPT-4 summaries are comparable to existing human-written impressions. The authors also collaborate with a board-certified radiologist to conduct a manual evaluation of the GPT-4 output. They state that *GPT-4 has a sufficient level of radiology knowledge with only occasional errors in complex context that require nuanced domain knowledge*[11]

Jain et al. conduct a survey of datasets and methods for clinical text summarization. The authors outline two major challenges to radiology report summarization that are common in the literature. First, in the clinical context, there is no room for factual inconsistencies or hallucinations[12] (a known limitation of LLMs). Secondly, the specific medical terminology found in the reports tends to be underrepresented in the datasets used to train LLMs for the general domain. This requires incorporating external medical knowledge bases.

Group relative policy optimization (GRPO) is a reinforcement learning (RL) algorithm designed by the authors of DeepSeekMath[4] for improving mathematical reasoning in LLMs. GRPO generates a group of candidate responses for each prompt, scores each with a reward model, and uses the group's average reward as a baseline to compute an advantage for each response when updating the policy. Unlike other RL algorithms, such as Proximal Policy Optimization[13], GRPO does not require a dedicated critic model which makes it very resource-efficient. The applications of GRPO for clinical natural language processing remain relatively unexplored. The authors of [14] use GRPO as the base reinforcement learning algorithm in their multiagent framework for multimodal medical reasoning which emulates a structured clinical workflow. The framework includes a General Practitioner agent, which first triages the user question to a specific department (e.g. surgery, or radiology). Then, specialist models provide preliminary judgments on the question. Finally, the specialist responses are routed to another GP agent (the attending physician), which formulates the final response. GRPO is used to improve the reasoning of the two GP agents. It is also part of the proposed Curriculum-Based Multi-Agent Reinforcement Learning (C-MARL) algorithm used to train the attending physician agent to better understand specialist agent responses. The framework is reported to achieve an average performance gain of 20.7% over supervised fine-tuning baselines[14]. Furthermore, an ablation study demonstrates that the GRPO-based C-MARL algorithm improves the capability of the attending physician agent to understand knowledge provided by the specialized agents by 15.7%.

3. Data

The MultiClinSum gold Spanish dataset[1] features 3998 examples of full-text clinical case reports and their corresponding human-generated summaries. The dataset authors provide 592 examples for training, with the remaining 3406 comprising the official test set. We further split the train set, leaving 500 examples for training and 92 for validation.

The clinical case reports are unstructured documents extracted from open journals and cover various specialties. Each report describes the medical history of a patient, their symptoms, tests, findings, diagnosis and treatment. Given the full text of a clinical case report in Spanish, our system is required to generate a summary which faithfully captures key information from the original report.

On average, case reports in the train set contain 1163 tokens (Llama-3.1-8B-Instruct tokenizer), with the longest document having 5486 tokens. The average number of tokens in the summaries is 217, and the max is 629. Consequently, our solution must be able to handle longer sequence lengths, so traditional summarization models, such as FLAN-T5[15] (with context length of 512 tokens), are less favorable or would require a sliding-window-based approach.

4. Methods

Large language models have shown promising results in clinical text summarization. The authors in [7] show that, when adapted on medical corpora, these models are capable of generating summaries that are often preferable to human summaries in terms of completeness and correctness. Thus, for the task of clinical case report summarization in Spanish, we also experiment with different approaches to adapting large language models.

All of our systems are centered around a single large language model which takes the full text of the report as input and produces a summary. The only postprocessing step is trimming any leading and trailing white spaces. Figure 1 illustrates the summarization flow.

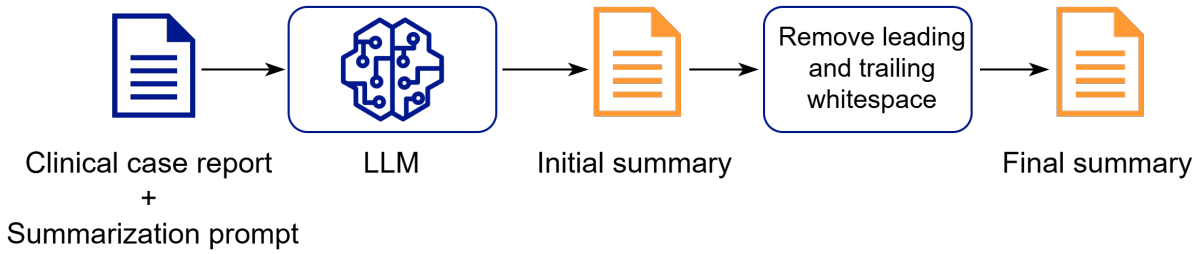

Figure 1: End-to-end clinical case report summarization flow - the full text of the clinical case report is passed to a large language model, along with summarization instructions. Then, the model generates a summary. Finally, any leading or trailing whitespace is removed from the summary text, producing the final summary.

4.1. Language Model Selection

We experiment with the following language models:

1. **Llama 3.1 8B Instruct** [3] - an open-source multilingual (including Spanish) auto-regressive large language model with 8 billion parameters. It is pre-trained on 15T multilingual tokens from the general domain and then instruction-tuned via supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF). Its multilingual capabilities, large context window (128K tokens), and mid-range parameter count make it an excellent candidate for fine-tuning on consumer hardware. Also, the performance of Llama 3.1 8B Instruct in a 0-shot setting serves as a good baseline for our experiments.
2. **GPT-4.1**[16] - at the time of writing, the flagship model in the GPT series by OpenAI. This proprietary model has a context window of 1,047,576 tokens and can generate texts in Spanish and other languages. We evaluate its 0-shot summarization capabilities to understand how a general LLM, that is readily accessible through a public API, compares to specialized models.
3. **Bio-Medical Llama 3 8B**[17] - a Llama 3 8B Instruct model further fine-tuned on a proprietary biomedical dataset, comprising of synthetic and manually curated samples. We also evaluate it in a 0-shot setting in an attempt to assess the impact of domain fine-tuning on summarization performance.

Model selection is primarily driven by resource constraints — not only in our experiment setup, but also in potential production scenarios where models may be deployed within clinical facilities rather than a datacenter. Also, we choose models that can work on Spanish texts.

4.2. Supervised Fine-tuning

We fine-tune the Unsloth[18] checkpoint of the Llama 3.1 8B Instruct model using the supervised fine-tuning trainer from Huggingface[19] on the train set of 500 report-summary pairs for 1 epoch. The complete list of hyperparameter values can be found in the provided source code. Most of the values are the defaults recommended by the Unsloth library, with the max sequence length and batch size adjusted for the particular dataset and the available compute resources.

4.3. Reinforcement Learning

The next class of methods we experiment with is reinforcement learning (RL), and more specifically, Group Relative Policy Optimization (GRPO)[4].

GRPO has been shown to boost LLM performance in various domains, including mathematical reasoning[4] and multimodal medical reasoning[14]. Unlike standard fine-tuning, which relies only on cross-entropy loss, GRPO allows us to optimize the BERTScore F1 and RougeL F1 metrics more directly by defining reward functions for each.

We choose GRPO as our base reinforcement learning algorithm because:

- GRPO is lightweight, requiring fewer computational resources than other reinforcement learning methods, such as Proximal Policy Optimization (PPO). Unlike PPO, GRPO does not rely on a separate critic model, which is typically the size of the policy model (the LLM) and must also be updated during training. This allows us to run our experiments on consumer hardware in a reasonable time.
- GRPO is well-suited for LLM training where a reward is assigned once the answer sequence is complete (i.e. it is assigned to the last token[4]). To distinguish good responses, GRPO uses the average reward of multiple sampled responses to the same question as a baseline[4]. As a result, unlike PPO, there is no need to approximate a value function during training for each token (i.e., to train a critic model). This is advantageous because, in the context of clinical report summarization, it only makes sense to assign a value to the complete summary rather than to intermediate partially generated sequences.
- The applicability of GRPO to clinical NLP tasks remains largely unexplored. To the best of our knowledge, no prior work has investigated the use of GRPO for clinical case report summarization.

We compare two training setups for Llama 3.1 8B Instruct on the 500 examples train set. Both setups share the same hyperparameter values and train for 1 epoch, with 6 candidate summaries generated per example. The main differences between the setups are in the system prompt, and the reward functions:

1. **GRPO Llama 3.1 Summarization** - this setup uses a system prompt which instructs the model to summarize the provided clinical case report, and preserve relevant clinical information, diagnosis, interventions, outcomes, and other fundamental aspects of the case. It also asks for the summary to be wrapped in <summary></summary> tags. The prompts can be found in the provided code. The setup features a reward function with a weight of 0.5 to incentivize the model to follow the expected response format. There is also a function that rewards the model 1.5 * RougeL F1, and a third function giving 3 * BERTScore F1. Naturally, these reward rules put a heavy emphasis on the BERTScore F1, since it is a semantic similarity metric.
2. **GRPO Llama 3.1 Planing and Summarization** - expands the previous Summarization setup by introducing a <plan-and-thoughts> output section prior to the <summary> section. The intuition here is to incentivize the model to create a planning/reasoning trace before generating a summary, so some of the tokens in that trace might help improve the generated summary by, for instance, preserving some key information from the case report. For this setup, there a four reward functions. The first one grants 0.125 points if the output contains a tag (<summary>, <plan-and-thoughts> or the corresponding closing tags). There is another function to award 0.5 points per correct pair of tags, further reinforcing the expected output format. Finally, we have the two functions for BERTScore F1 and ROUGE-L F1. What differs from the previous setup is

that the weight of BERTScore F1 is set to 4, to compensate for the increase in weight from the formatting reward functions. This way, we ensure that BERTScore F1 has the highest impact on the total reward of all metrics.

We use the Llama 3.1 8B instruct model instead of the Bio-Medical Llama 3 8B for the GRPO training due to a technical difficulty with the training script. It was resolved after the task deadline, and the result is shown in table 2. The *GRPO Bio-Medical Llama 3 Summarization* model is trained using the exact same setup as in *GRPO Llama 3.1 Summarization*.

5. Experiments and Results

In this section, we analyze the performance of the different model adaptation approaches. Table 2 shows the average ROUGE-L and BERTScore results on the validation set.

5.1. Evaluation Metrics

The official evaluation metrics are ROUGE-L and BERTScore. In our experiments we use bert-base-multilingual-cased[20] for calculating BERTScore.

ROUGE-L[5] is based on the longest common subsequence of words between a generated sequence and a reference sequence. It measures the longest sequence of words that appear in the same order in both candidate and reference summaries, even if the words are not contiguous. This allows it to capture sequence-level syntactic similarity without requiring exact matches.

BERTScore[6] computes the semantic similarity between the candidate and reference summaries. It uses a BERT-based model [20] to obtain contextualized representations of the tokens in both sequences, and then calculates pairwise cosine similarity.

5.2. Hardware Setup

All experiments were conducted in a Google Collab Pro environment. Two setups were used depending on the requirements of the particular language model:

1. **Single NVIDIA L4 GPU (22.5 GB VRAM) + 53 GB RAM** - for the supervised fine-tuning of Llama 3.1 8B Instruct, as well as inference.
2. **Single NVIDIA A100 GPU (40 GB VRAM) + 83.5 GB RAM** - for the GRPO training of Llama 3.1 8B Instruct.

5.3. Baseline

We use the unaltered Llama 3.1 8B-Instruct, and the gpt-4.1-2025-04-14 models as a baseline. We simply provide a system prompt, followed by a user prompt containing the case report text to summarize, with no examples. The system prompt describes the model role (*a physician's AI assistant*), the task (*summarizing clinical case reports*), and provides instructions about the tone and format (*clear, precise, and coherent summary*). The prompts can be found in the provided code.

The unmodified Llama 3.1 8B-Instruct shows strong BERTScore results, outperforming GPT-4.1 in terms of ROUGE-L and BERTScore precision. On the other hand, GPT-4.1 has higher ROUGE-L recall and F1 than Llama, indicating that its summaries align more closely syntactically with the reference summaries, and likely capture a greater portion of the key information.

5.4. Effect of Biomedical Pre-training

The Bio-Medical Llama 3 8B performs similarly to the unmodified Llama 3.1 8B, with a slight improvement in BERTScore precision. This, combined with comparable ROUGE-L precision scores, suggests that the summaries generated by the biomedical model align more closely with the reference summaries in terms of the information they convey, although it may be rephrased.

5.5. Effect of Fine-tuning

Fine-tuning the Llama 3.1 8B Instruct model greatly improves ROUGE-L recall compared to the unaltered model, which indicates that summaries generated by it feature more of the phrasing from the reference summaries. However, the significant drop in the other metrics suggests that the training set may be too small or that the number training epochs may be insufficient.

5.6. Effect of GRPO

Of all models, *GRPO Llama 3.1 Summarization* performs best in terms of both ROUGE-L F1 and BERTScore F1 on the validation set. There is a stable increase in ROUGE-L and BERTScore precision, showing that the generated summaries more closely match the reference ones, both syntactically and semantically. The model also consistently follows the expected response format.

We use this model for our single submission run on the test set. Test set results are shown in table 1.

Table 1
Average ROUGE-L and BERTScore results on the test set.

Model	ROUGE-L Precision	ROUGE-L Recall	ROUGE-L F1	BERTScore Precision	BERTScore Recall	BERTScore F1
GRPO Llama 3.1 Summarization	0.3639	0.2667	0.2899	0.7699	0.7470	0.7578

Surprisingly, the *GRPO Llama 3.1 Planing and Summarization* model falls behind the baseline, despite showing a stable increase in ROUGE-L recall. Perhaps, the additional reward functions introduce competing objectives during training, shifting the focus from the ROUGE-L and BERTScore metrics.

When it comes to the *GRPO Bio-Medical Llama 3 Summarization* model, we observe a stable increase in ROUGE-L scores compared to the base Bio-Medical Llama 3 8B. This shows that the GRPO model produces summaries that better align with the reference summaries in terms of sequence and phrasing, possibly suggesting an improved retention of medical phrases. Comparable BERTScore F1 scores indicate that the GRPO training does not negatively impact the model's ability to produce summaries that are semantically similar to the reference ones. Finally, we make an interesting observation during GRPO training - the Bio-Medical Llama model does not adhere to the required response format (<summary></summary>), although it produces coherent summaries. This might be because the proprietary BioMedData dataset used for training the model is in English, which may lead to reduced instruction-following capabilities in Spanish.

Table 2
Average ROUGE-L and BERTScore results on the validation set.

Model	ROUGE-L Precision	ROUGE-L Recall	ROUGE-L F1	BERTScore Precision	BERTScore Recall	BERTScore F1
Llama 3.1 8B	0.2207	0.4257	0.2804	0.7206	0.7794	0.7529
GPT-4.1	0.1729	**0.5169**	0.2592	0.69899	**0.7840**	0.7391
Bio-Medical Llama 3 8B	0.2240	0.3146	0.2618	0.7415	0.7742	0.7575
Fine-tuned Llama 3.1	0.1164	0.4944	0.1884	0.6769	0.7540	0.7133
GRPO Llama 3.1 Summarization	0.2636	0.3677	**0.2946**	**0.7466**	0.7712	**0.7583**
GRPO Llama 3.1 Planing and Summarization	0.1288	0.4719	0.2024	0.6880	0.7782	0.73038
GRPO Bio-Medical Llama 3 Summarization	**0.2679**	0.3474	0.2838	0.74163	0.76152	0.7508

6. Conclusion

We evaluate several approaches to adapting large language models for Spanish clinical case report summarization. Our best performing model is trained using GRPO with reward functions that reflect BERTScore, ROUGE-L and response formatting. It shows that reinforcement learning can be applied to an instruction-tuned model to increase summarization precision while maintaining good recall.

Future work could explore whether GRPO or any other RL algorithm can improve the performance of language models pre-trained on Spanish clinical corpora. Furthermore, since GRPO allows for a variety of reward functions, it would be interesting to experiment with, for instance, a small monolingual domain-specific evaluator model as a reward function. Finally, due to our limited compute budget, we only experimented with language models in the 8B parameter range. Exploring whether the proposed GRPO-based summarization approach scales well, by applying it to larger models, is another important direction for future work.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and Overleaf's Writefull in order to: Grammar and spelling check, Paraphrase and reword. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] M. Rodríguez-Ortega, E. Rodríguez-Lopez, S. Lima-López, C. Escolano, M. Melero, L. Pratesi, L. Vigil-Gimenez, L. Fernandez, E. Farré-Maduell, M. Krallinger, Overview of MultiClinSum task at BioASQ 2025: evaluation of clinical case summarization strategies for multiple languages: data, evaluation, resources and results., in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[2] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, F. N. Maria Di Nunzio, Giorgio, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, in: L. P. A. G. S. d. H. J. M. F. P. P. R. D. S. G. F. N. F. Jorge Carrillo-de Albornoz, Julio Gonzalo (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[3] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[4] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, D. Guo, Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL: https://arxiv.org/abs/2402.03300. arXiv:2402.03300.

[5] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[6] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2020. URL: https://arxiv.org/abs/1904.09675. arXiv:1904.09675.

[7] D. Van Veen, C. Van Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly, A. S. Chaudhari, Clinical text summarization: Adapting large language models can outperform human experts, Res Sq (2023).

[8] Z. Huemann, C. Lee, J. Hu, S. Y. Cho, T. Bradshaw, Domain-adapted large language models for classifying nuclear medicine reports, 2023. URL: https://arxiv.org/abs/2303.01258. arXiv:2303.01258.

[9] H. Y. Leong, Y. F. Gao, Shuai Ji, Uktu Pamuksuz, Efficient fine-tuning of large language models for automated medical documentation (2024). URL: https://rgdoi.net/10.13140/RG.2.2.26884.74881. doi:10.13140/RG.2.2.26884.74881.

[10] W. wai Yim, Y. Fu, A. B. Abacha, N. Snider, T. Lin, M. Yetisgen, Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation, 2023. URL: https://arxiv.org/abs/2306.02022. arXiv:2306.02022.

[11] Q. Liu, S. Hyland, S. Bannur, K. Bouzid, D. C. Castro, M. T. Wetscherek, R. Tinn, H. Sharma, F. Pérez-García, A. Schwaighofer, P. Rajpurkar, S. T. Khanna, H. Poon, N. Usuyama, A. Thieme, A. V. Nori, M. P. Lungren, O. Oktay, J. Alvarez-Valle, Exploring the boundaries of gpt-4 in radiology, 2023. URL: https://arxiv.org/abs/2310.14573. arXiv:2310.14573.

[12] R. Jain, A. Jangra, S. Saha, A. Jatowt, A survey on medical document summarization, 2022. URL: https://arxiv.org/abs/2212.01669. arXiv:2212.01669.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, 2017. URL: https://arxiv.org/abs/1707.06347. arXiv:1707.06347.

[14] P. Xia, J. Wang, Y. Peng, K. Zeng, X. Wu, X. Tang, H. Zhu, Y. Li, S. Liu, Y. Lu, et al., Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning, arXiv preprint arXiv:2506.00555 (2025).

[15] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, et al., Scaling instruction-finetuned language models, 2022. URL: https://arxiv.org/abs/2210.11416. arXiv:2210.11416.

[16] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, et al., Gpt-4 technical report, 2024. URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[17] Contactdoctor-bio-medical: A high-performance biomedical language model, https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B, 2024.

[18] M. H. Daniel Han, U. team, Unsloth, 2023. URL: http://github.com/unslothai/unsloth.

[19] L. von Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert, S. Huang, K. Rasul, Q. Gallouédec, Trl: Transformer reinforcement learning, https://github.com/huggingface/trl, 2020.

[20] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.