Supervised Fine-tuning of DeepSeek-R1-Qwen-1.5b

Thomas Cotter
thomas.cotter3@dxc.com

February 26, 2025

1 Abstract

In this mini-report, we discuss the process of using Supervised Fine-tuning (SFT) to produce a Low Rank Adapter (LoRA) that guides the model into producing a "Final Answer" after its reasoning and initial output. The capability enhances the model's usability by providing clear, concise conclusions following its chain-of-thought reasoning. Using our in-house Synthetic Data Generator, we modified an open-source medical reasoning dataset to include a single sentence summary of the answer.

We successfully implemented this fine-tuning on a single NVidia A100 GPU, demonstrating the feasibility of adapting foundational models with limited computational resources. This work has potential in many "knowledge worker" use cases, where both a transparent reasoning process and a concise, actionable answer are required, particularly in medical contexts where decision support must balance thoroughness with clarity.

2 Introduction

Given the success of DeepSeek R1 [3], along with the addition of it as a highly capable reasoning model to the ever-growing number of open-source models, we wanted to determine the viability of adapting a distilled version of R1 to respond with a different form, without losing the quality that comes with reasoning.

Previous work on SFT from DeepSeek was adapting a non-reasoning model (Qwen2.5) to be a reasoning model by training it on 800K outputs of the base R1 model. Other work on fine-tuning reasoning models uses other techniques like Reinforcement Fine-tuning [4]. However, this is useful when you have highly verifiable outputs, such as producing code which can then be run against unit tests. For our case, adapting the model to produce similar, but structured outputs SFT is more suited.

3 Methodology

3.1 Dataset

Initially, we wanted to adapt the model to an internal dataset. However, since the base model is a reasoning model, we must use a reasoning dataset. Without one, applying SFT would guide the model away from reasoning, and more towards a generic LLM answer.

One potential mitigation to this problem would be to ignore the effect of any generated reasoning tokens on the overall loss. Unsloth [2] (a fine-tuning library) implement this well in `train_on_responses_only()` - ignoring the effect of the user question on the loss, and this could be extended to all reasoning tokens. We leave this to future work to determine the viability of this.

Given that we were required a reasoning dataset, we needed to generate our dataset from a reasoning model, similar to how DeepSeek produced the set of distilled models initially. Fortunately, a freely available reasoning dataset exists on HuggingFace[1], created by Chen et al. [1]. This dataset was created from a series of medical benchmarks. Initially, OpenAIs GPT-4o [6] is repeatedly prompted using a variety of techniques until the model gets the answer correct. The entire search for the correct answer is compressed and formatted into a single complex CoT reasoning process, which means we obtain both a full reasoning process and a final correct answer.

We extended this dataset using our in-house Synthetic Data Generator to add an additional aspect to the response, a *final summary*, prompting Claude Sonnet 3.5 (new) to do so. Given all of this, we could construct examples to provide to our model that look like Figure 1.

[1] https://huggingface.co/datasets/FreedomIntelligence/medical-o1-reasoning-SFT

{question}
<think>
{reasoning}
</think>
{output}
Final Answer: {summary}

Figure 1: The style of input provided to the model during fine-tuning

By providing examples in this way, not only would our LoRA guide the model into answering in a way more similar to the outputs collated by GPT-4o, but also always respond with a `Final Answer:` at the end. Our final dataset is publically released on HuggingFace[2].

3.2 Fine-tuning

In order to finetune our model, we used Unsloth [2]. Unsloth is a fine-tuning library that reduces the time & memory requirements for fine-tuning compared to Transformers. Their team have created a series of 4bit quantized version of popular open source models, including the model we want to finetune: `https://huggingface.co/unsloth/DeepSeek-R1-Distill-Qwen-1.5B-unsloth-bnb-4bit`. Fine-tuning this quantized version allows for much faster iteration speeds.

For the fine-tuning, we created a LoRA with a rank of 16 and an alpha of 32. This ratio is recommended by Shuttleworth et al. [8]. They also recommend to use rank-stabilized LoRA [5], which adjusts the scaling factor of the adapters. We included this as well.

In terms of hyperparameters, we used the Adam 8bit optimizer, a learning rate of 2e-4 and an effective batch size of 128. We trained the model for 100 steps on 23000 samples.

4 Results

Given our test set of 326 questions (not included in the fine-tuning dataset), we compared the base model to our finetuned model on two metrics: **(1) Presence of 'Final Answer'** and **(2) Accuracy**. We expected that **(1)** should significantly increase after fine-tuning, since this was our goal, and we hope that **(2)** does not decrease. This is because LoRA SFT is not adding knowledge to the model, it's simply guiding it to produce a different style of output. We would not expect the model to improve on medical questions it has not seen before. Our results for this experiement (tested against 326 unseen questions) can be seen in Table 1.

Model Name	Presence of 'Final Answer:'	Accuracy
R1-Distilled-Qwen1.5B (Base)	0.043	**0.181**
R1-Distilled-Qwen1.5B (Finetuned)	**0.997**	0.132

Table 1: Comparison of `DeepSeek-R1-Distill-Qwen-1.5B-unsloth-bnb-4bit` finetune

To measure accuracy, we used LLM-as-a-Judge [9], with Claude Haiku 3.5 acting as judge. Since we know that the responses in the original Chen et al. dataset are correct, we can ask the judge to compare those to our generated responses.

To determine if the difference in accuracy between the base and fine-tuned models was statistically significant, we applied McNemar's test, which is appropriate for comparing paired nominal data. The test examines the symmetry of disagreements between the two models. We obtained a p-value of 0.039, which is below the conventional significance threshold of 0.05, indicating that the accuracy decrease is statistically significant rather than due to random chance. This accuracy loss was somewhat expected, as we fine-tuned the R1-based model using a dataset derived from GPT-4o's reasoning patterns. Since GPT-4o has been shown to underperform R1 on reasoning tasks, adapting our model to emulate GPT-4o's reasoning structure likely contributed to the observed decrease in accuracy.

5 Conclusion

A concise final summary after reasoning is valuable as it reduces cognitive load for users who need quick answers but can reference reasoning when needed. In wider terms, we also showed that it's possible to adapt a reasoning model to have a structured output, which becomes important when adding these models to agentic systems. Agentic systems will favour smaller models due to reduced computational requirements, and we proved here that it's possible to format the outputs

[2] `https://huggingface.co/datasets/tcotter/o1-medical-data-reasoning-w-summary`

of a small (1.5B) model without prompt engineering. This enhances the practical utility of reasoning models in real-world applications.

Our work demonstrates an effective fine-tuning approach for modifying output format without prompt engineering. This shows promise for building AI systems to help users who require both a short summary for the next actions, along with a long reasoning process for review if required. This approach is especially valuable in high-stakes domains like medicine, finance, and legal applications, where both efficiency and reasoning transparency are critical for responsible deployment and human oversight. As smaller reasoning models continue to improve, techniques that enhance their practical utility while preserving their reasoning capabilities will play an increasingly important role in the responsible advancement of AI systems.

5.1 Future Work

Whilst we did see a significant improvement in the presence of a "Final Answer", the accuracy did drop slightly. This shows that more work needs to be done in being able to adapt these models to produce formatted outputs, whilst still keeping the high level of performance. Recently, Chen et al. have released an additional dataset[3], created by DeepSeekR1 rather than GPT-4o. Using this dataset might prove to produce better results, since we know that R1 performs better than GPT-4o. Other future work will include using GRPO [7] to perform reinforcement fine-tuning or by adapting the training process to ignore the effect of reasoning tokens on the loss. Both of these techniques require further research, though reinforcement fine-tuning appears most promising as it could directly optimize for maintaining accuracy while enforcing the desired output format.

References

[1] Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms, 2024.

[2] Michael Han Daniel Han and Unsloth team. Unsloth, 2023.

[3] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[4] Arnav Garg, Travis Addair, and Will Van Eaton. Teaching ai to write gpu code: A deep dive into reinforcement fine-tuning, 2025.

[5] Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora, 2023.

[6] OpenAI. Gpt-4o system card, 2024. Accessed: 2025-02-25.

[7] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

[3] https://huggingface.co/datasets/FreedomIntelligence/Medical-R1-Distill-Data

[8] Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence, 2024.

[9] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.