CoT Distillation LLMs enable cost effective Artificial Intelligence for On-Premises Services

Why I chose this subject

I am interested in creating an actual proposal to some co-workers on this subject. The idea of presenting a cost-effective tool that skirts some of the legal challenges vendor provided AI as a service have seems quite attractive. Additionally, I would develop a greater understanding of LLMs and how they might be beneficial to my industry.

Synopsis

Given the recent industry rocking discoveries in DeepSeek V3 and other distillation LLMs and costs associated with AI as a service, On-Premises AI services begin to gain traction as an attractive solution for small to medium setups. Beginning with other off-the-shelf tools such as vLLMs, the idea of On-Premises AI services has materialized into experimental projects investigating the viability of internal AI tooling. This paper will explore how CoT Distillation can produce models with greater portability, cost efficiency, and capability compared to alternative open-source models such as vLLMs.

Outline

1. Introduction

The problem I have solved

- Main Assertion: Present a POC that shows notable improvement from vLLMs to Distilled models, which unlock performant on-premises and on the edge deployments with real world applications. LiteLLM allows for model agnostic communication, allowing for easy model changes.

Why the problem is not already solved or other solutions are ineffective in one or more important ways

- Due to the rapidly changing nature of AI, it is difficult to have a successful deployment of AI tools.
- The new developments with COT Distilled models
- The current landscape of AI tooling for the enterprise – Amazon Bedrock
- Cost/maintenance issues associated with On-Premises services

Why my solution is worth considering and why is it effective in some way that others are not

- By investing in the technical resourcing, it will unlock new innovative and cost effective solutioning not only for on the edge applications but will more importantly empower greater decision making in the entire enterprise space. On premises solutions enable faster more efficient improvements and unlock unique applications.

How the rest of the paper is structured

- The rest of this paper first discusses related work in Section 2, and then describes my implementation in Section 3. Section 4 describes how I evaluated my system and presents the results. Section 5 presents my conclusions and describes future work.

2. Related Work

Other efforts that exist to solve this problem and why are they less effective than my method

- “On-Premise Artificial Intelligence as a Service” paper does not address security and legal concerns with vendor provided AI tooling
- The models presented in the COT Distillation papers are helpful but provide little application in this specific field

Other efforts that exist to solve related problems that are relevant, how are they relevant, and why are they less effective than my solution for this problem

- The costs associated with Amazon Outposts is prohibitive for effective innovation

3. Implementation

What I Did: *My Solution*

- Tested two on-edge applications that use the new DeepSeek distilled model with a RAG provided with help docs on Disneyland Parks:
    - IOS Application
    - Linux based python script with TTS

How my solution works

- Pending Implementation

4. Evaluation

How I tested my solution

- Performance metrics
    - Character per minuets
    - Power usage
    - CPU usage
    - VRAM requirements and usage
- Platforms
    - Desktop
    - Mobile application
    - Integrated Systems?

How my solution performed, how its performance compared to that of other solutions mentioned in related work, and how these results show that my solution is effective

- Pending Implementation
- What the results *do* and *do not* say

5. Conclusions and Future Work

- Reiterate:
    - the problem I have solved
    - My solution to the problem
    - Why my solution is worthwhile in some significant way
- What I could do next
    - Improve my solution
    - Apply my solution to harder or more realistic versions of this problem
    - Apply my solution or a related solution to a related problem

Works Cited

Related Journals

- DeepSeek-V3 Technical Report
- Learning to Maximize Mutual Information for Chain-of-Thought Distillation
- On-Premise Artificial Intelligence as a Service for Small and Medium Size Setups

Repositories

- https://github.com/BerriAI/litellm
- https://github.com/weaviate/weaviate
- https://github.com/vllm-project/vllm
- https://docs.aws.amazon.com/pdfs/bedrock/latest/userguide/bedrock-ug.pdf#what-is-bedrock
- https://docs.aws.amazon.com/outposts/latest/userguide/what-is-outposts.html

Primary

Choi, W., Kim, W. K., Yoo, M., & Woo, H. (2024). Embodied cot distillation from llm to off-the-shelf agents. https://arxiv.org/abs/2412.11499

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., . . . Pan, Z. (2024). Deepseek-v3 technical report. https://arxiv.org/abs/2412.19437

Fortuna, C., Muši´c, D., Cerar, G., ˇCampa, A., Kapsalis, P., & Mohorˇciˇc, M. (2022). On-premise artificial intelligence as a service for small and medium size setups. https://arxiv.org/abs/2210.06956

Secondary

Gloeckle, F., Idrissi, B. Y., Rozière, B., Lopez-Paz, D., & Synnaeve, G. (2024). Better faster large language models via multi-token prediction. https://arxiv.org/abs/2404.19737

Peng, H., Wu, K., Wei, Y., Zhao, G., Yang, Y., Liu, Z., Xiong, Y., Yang, Z., Ni, B., Hu, J., Li, R., Zhang, M., Li, C., Ning, J., Wang, R., Zhang, Z., Liu, S., Chau, J., Hu, H., & Cheng, P. (2023). Fp8-lm: Training fp8 large language models. https://arxiv.org/abs/2310.18313

Source BibTeX Code

```
@misc{choi2024embodiedcotdistillationllm,
      title={Embodied CoT Distillation From LLM To Off-the-shelf Agents},
      author={Wonje Choi and Woo Kyung Kim and Minjong Yoo and Honguk Woo},
      year={2024},
      eprint={2412.11499},
      archivePrefix={arXiv},
      primaryClass={cs.AI},
      url={https://arxiv.org/abs/2412.11499},
}

@misc{deepseekai2024deepseekv3technicalreport,
      title={DeepSeek-V3 Technical Report},
      author={DeepSeek-AI and Aixin Liu and Bei Feng and Bing Xue and Bingxuan Wang and Bochao Wu and Chengda Lu
and Chenggang Zhao and Chengqi Deng and Chenyu Zhang and Chong Ruan and Damai Dai and Daya Guo and Dejian Yang and
Deli Chen and Dongjie Ji and Erhang Li and Fangyun Lin and Fucong Dai and Fuli Luo and Guangbo Hao and Guanting
Chen and Guowei Li and H. Zhang and Han Bao and Hanwei Xu and Haocheng Wang and Haowei Zhang and Honghui Ding and
Huajian Xin and Huazuo Gao and Hui Li and Hui Qu and J. L. Cai and Jian Liang and Jianzhong Guo and Jiaqi Ni and
Jiashi Li and Jiawei Wang and Jin Chen and Jingchang Chen and Jingyang Yuan and Junjie Qiu and Junlong Li and
Junxiao Song and Kai Dong and Kai Hu and Kaige Gao and Kang Guan and Kexin Huang and Kuai Yu and Lean Wang and
Lecong Zhang and Lei Xu and Leyi Xia and Liang Zhao and Litong Wang and Liyue Zhang and Meng Li and Miaojun Wang
and Mingchuan Zhang and Minghua Zhang and Minghui Tang and Mingming Li and Ning Tian and Panpan Huang and Peiyi
Wang and Peng Zhang and Qiancheng Wang and Qihao Zhu and Qinyu Chen and Qiushi Du and R. J. Chen and R. L. Jin and
Ruiqi Ge and Ruisong Zhang and Ruizhe Pan and Runji Wang and Runxin Xu and Ruoyu Zhang and Ruyi Chen and S. S. Li
and Shanghao Lu and Shangyan Zhou and Shanhuang Chen and Shaoqing Wu and Shengfeng Ye and Shengfeng Ye and Shirong
Ma and Shiyu Wang and Shuang Zhou and Shuiping Yu and Shunfeng Zhou and Shuting Pan and T. Wang and Tao Yun and
Tian Pei and Tianyu Sun and W. L. Xiao and Wangding Zeng and Wanjia Zhao and Wei An and Wen Liu and Wenfeng Liang
and Wenjun Gao and Wenqin Yu and Wentao Zhang and X. Q. Li and Xiangyue Jin and Xianzu Wang and Xiao Bi and
Xiaodong Liu and Xiaohan Wang and Xiaojin Shen and Xiaokang Chen and Xiaokang Zhang and Xiaosha Chen and Xiaotao
Nie and Xiaowen Sun and Xiaoxiang Wang and Xin Cheng and Xin Liu and Xin Xie and Xingchao Liu and Xingkai Yu and
Xinnan Song and Xinxia Shan and Xinyi Zhou and Xinyu Yang and Xinyuan Li and Xuecheng Su and Xuheng Lin and Y. K.
Li and Y. Q. Wang and Y. X. Wei and Y. X. Zhu and Yang Zhang and Yanhong Xu and Yanhong Xu and Yanping Huang and
Yao Li and Yao Zhao and Yaofeng Sun and Yaohui Li and Yaohui Wang and Yi Yu and Yi Zheng and Yichao Zhang and Yifan
Shi and Yiliang Xiong and Ying He and Ying Tang and Yishi Piao and Yisong Wang and Yixuan Tan and Yiyang Ma and
Yiyuan Liu and Yongqiang Guo and Yu Wu and Yuan Ou and Yuchen Zhu and Yuduan Wang and Yue Gong and Yuheng Zou and
Yujia He and Yukun Zha and Yunfan Xiong and Yunxian Ma and Yuting Yan and Yuxiang Luo and Yuxiang You and Yuxuan
Liu and Yuyang Zhou and Z. F. Wu and Z. Z. Ren and Zehui Ren and Zhangli Sha and Zhe Fu and Zhean Xu and Zhen Huang
and Zhen Zhang and Zhenda Xie and Zhengyan Zhang and Zhewen Hao and Zhibin Gou and Zhicheng Ma and Zhigang Yan and
Zhihong Shao and Zhipeng Xu and Zhiyu Wu and Zhongyu Zhang and Zhuoshu Li and Zihui Gu and Zijia Zhu and Zijun Liu
and Zilin Li and Ziwei Xie and Ziyang Song and Ziyi Gao and Zizheng Pan},
      year={2024},
      eprint={2412.19437},
      archivePrefix={arXiv},
      primaryClass={cs.CL},
      url={https://arxiv.org/abs/2412.19437},
}

@misc{fortuna2022onpremiseartificialintelligenceservice,
      title={On-Premise Artificial Intelligence as a Service for Small and Medium Size Setups},
      author={Carolina Fortuna and Din Mu≈°ifá and Gregor Cerar and Andrej fåampa and Panagiotis Kapsalis and
Mihael Mohorƒçifç},
      year={2022},
      eprint={2210.06956},
      archivePrefix={arXiv},
      primaryClass={cs.SE},
      url={https://arxiv.org/abs/2210.06956},
}

@misc{gloeckle2024betterfasterlarge,
      title={Better & Faster Large Language Models via Multi-token Prediction},
      author={Fabian Gloeckle and Badr Youbi Idrissi and Baptiste Rozi√®re and David Lopez-Paz and Gabriel
Synnaeve},
      year={2024},
      eprint={2404.19737},
      archivePrefix={arXiv},
      primaryClass={cs.CL},
      url={https://arxiv.org/abs/2404.19737},
}

@misc{peng2023fp8lmtrainingfp8large,
      title={FP8-LM: Training FP8 Large Language Models},
      author={Houwen Peng and Kan Wu and Yixuan Wei and Guoshuai Zhao and Yuxiang Yang and Ze Liu and Yifan Xiong
and Ziyue Yang and Bolin Ni and Jingcheng Hu and Ruihang Li and Miaosen Zhang and Chen Li and Jia Ning and Ruizhe
Wang and Zheng Zhang and Shuguang Liu and Joe Chau and Han Hu and Peng Cheng},
      year={2023},
      eprint={2310.18313},
      archivePrefix={arXiv},
      primaryClass={cs.LG},
      url={https://arxiv.org/abs/2310.18313},
}
```