VetMedGPT: Generative Pre-trained Transformer for Enhanced Animal Healthcare

Pinxue Lin, Sayed Raheel Hussain and Thirupathi Kadari,
M.S. in Artificial Intelligence

Faculty Advisor: Youshan Zhang, Ph.D.

Katz School of Science and Health

ABSTRACT

With the development of artificial intelligence (AI), language models have gradually matured, but general large language models struggle to cover all knowledge areas within the specialized field of veterinary medicine. This project introduces VetMedGPT, an innovative language model designed to enhance performance in the veterinary field through AI. Developed based on over 500GB of veterinary corpus, VetMedGPT aims to address critical knowledge gaps in animal diseases, treatments, and clinical procedures. This AI model was trained on veterinary-related datasets to strengthen the widespread dissemination of veterinary knowledge among the public by reducing deployment costs and maintaining acceptable accuracy. Its applications range from serving as an interactive learning tool in classroom teaching to assisting non-professionals in understanding animal symptoms and seeking medical help in a timely manner. The integration of VetMedGPT into veterinary education and research lays the foundation for transforming diagnostic and treatment patterns for animal diseases.

INTRODUCTION

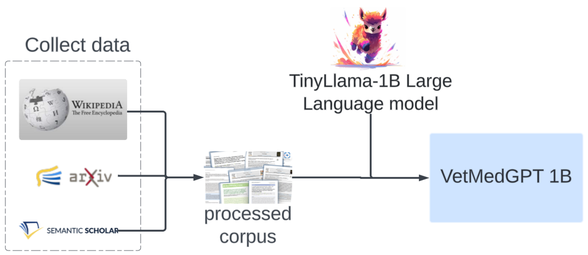
- The field of veterinary medicine stands on the cusp of transformation, with AI offering exciting possibilities.
- VetMedGPT is an innovative tool specifically designed for veterinary practice.
- VetMedGPT was created to address gaps in veterinary literature, particularly in understanding animal anatomy, diseases, treatments, and clinical procedures. This Generative Pre-trained Transformer is unique in its focus on animal physiology and terminology. It aims to simplify veterinary knowledge complexities and offer specialized support within veterinary sciences.
- By learning from 500GB of professional literature and datasets on animal physiology, VetMedGPT strives for accurate and relevant responses to veterinary questions. This promises to elevate its chat performance and ensure its usefulness in real-world situations.
- VetMedGPT represents a step toward improving clinical decision support, literature analysis, and veterinary education. It aims to bridge information gaps using advanced AI and has the potential to support veterinary care, research, and training, potentially improving efficiency and accuracy in animal healthcare.

METHODOLOGY

Data collection: The dataset comprised of 500GB of professional veterinary literature and normal animal physiological data. This data was sourced from peer-reviewed journals and Wikipedia entries to represent a wide array of veterinary knowledge domains.

Related tech: VetMedGPT was advanced using TinyLlama-1B (Zhang et al., 2024), a scaled-down variant of transformer models like ChatGPT (Vaswani et al., 2023). It has 1 billion parameters, far fewer than GPT-4's 175 billion. For efficiency, we applied technologies such as FSDP (Zhao et al., 2023) and Flash-Attention (Dao, 2023) during training.

Processing: The fine-tuning process began with an initial pre-processing phase, where the dataset was cleaned and structured to ensure consistency and relevance. Following this, the model underwent a training phase. After the initial training, it required further fine-tuning with question-and-answer pairs to enable conversational abilities.



RESULTS

Evaluate Task: The model was evaluated with standard question-answering and multiple-choice questions.

Evaluate Method: For multiple-choice questions, accuracy was used as the evaluation criterion (Table 1). For text questions, the rouge score between the generated answer and the standard answer was selected as the evaluation standard (Table 2). The rouge score can demonstrate the degree of correlation between two segments of text.

Table 1.
Result for Multiple-choice Questions

model	Accuracy
Llama2 7B chat	0.3173
TinyLlama 1B chat	0.1792
VetMedGPT 1B	0.2535
Mistral 7B Instruct-V0.2	0.4391
Falcon 7B Instruct	0.2368

Table 2.
Result for Text Question Answer Rouge Score

model	ROUGE-1			ROUGE-2			ROUGE-L		
	r	p	f	r	p	f	r	p	f
tinyllama 1B	0.3684	0.067	0.114	0.073	0.010	0.018	0.361	0.067	0.111
VetMedGPT 1B	0.391	0.073	0.122	0.093	0.012	0.021	0.356	0.066	0.110
llama2 7b chat	0.473	0.105	0.170	0.149	0.023	0.040	0.431	0.095	0.154
Falcon 7B	0.360	0.117	0.174	0.106	0.030	0.046	0.325	0.106	0.157
Mistral 7B Instruct-V0.2	0.491	0.095	0.158	0.143	0.020	0.035	0.451	0.087	0.145

Model Selection: We compared our 1B model with the original TinyLlama-1B model and included more 7B models (seven times larger than our model) to compare the results. The results show that our data has improved the performance of the original model in the field of veterinary medicine.

CONCLUSIONS

Findings: Our analysis indicates that data, when fine-tuned, yields answers with a stronger relevance in the field of veterinary medicine compared to those generated by the original model. This suggests a tailored approach enhances the model's applicability to specialized domains.

Limitations: Our smaller model has not reached the performance of larger models, especially in complex tasks requiring deep context.

Future Work:

1. Integrate RAG for improvement.
2. Training with larger models to achieve greater comprehension abilities and enhanced logical reasoning.

Summary: The development of VetMedGPT, a tailored language model finetuned for veterinary healthcare, fills a critical void in AI-supported animal healthcare. By harnessing specialized datasets and advanced training methods, we've demonstrated AI's potential to enhance both the accessibility and quality of animal care.

REFERENCES

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need. arXiv preprint. arXiv:1706.03762.

Zhao, Y., Gu, A., Varma, R., Luo, L., Huang, C.-C., Xu, M., Wright, L., Shojanazeri, H., Ott, M., Shleifer, S., ... Li, S. (2023). PyTorch FSDP: Experiences on scaling fully sharded data parallel. arXiv preprint arXiv:2304.11277.

Dao, T. (2023). FlashAttention-2: Faster attention with better parallelism and work partitioning. arXiv preprint arXiv:2307.08691.

Zhang, P., Zeng, G., Wang, T., & Lu, W. (2024). TinyLlama: An open-source small language model. arXiv preprint arXiv:2401.02385.