VetLLM: Large Language Model for Predicting Diagnosis from Veterinary Notes

Yixing Jiang, Jeremy A. Irvin, Andrew Y. Ng and James Zou†
Stanford University, Stanford, CA, United States †*E-mail: jamesz@stanford.edu*

Lack of diagnosis coding is a barrier to leveraging veterinary notes for medical and public health research. Previous work is limited to develop specialized rule-based or customized supervised learning models to predict diagnosis coding, which is tedious and not easily transferable. In this work, we show that open-source large language models (LLMs) pretrained on general corpus can achieve reasonable performance in a zero-shot setting. Alpaca-7B can achieve a zero-shot F1 of 0.538 on CSU test data and 0.389 on PP test data, two standard benchmarks for coding from veterinary notes. Furthermore, with appropriate fine-tuning, the performance of LLMs can be substantially boosted, exceeding those of strong state-of-the-art supervised models. VetLLM, which is fine-tuned on Alpaca-7B using just 5000 veterinary notes, can achieve a F1 of 0.747 on CSU test data and 0.637 on PP test data. It is of note that our fine-tuning is data-efficient: using 200 notes can outperform supervised models trained with more than 100,000 notes. The findings demonstrate the great potential of leveraging LLMs for language processing tasks in medicine, and we advocate this new paradigm for processing clinical text.

Keywords: Diagnosis Extraction, Veterinary Notes, Veterinary Medicine, Large Language Models, LLM, Foundation Models.

1. Introduction

Most veterinary records are in free-text forms without structured diagnostic codes, making it difficult to use for medical research, public health monitoring or quality-improvement programs.[1] For example, the eligibility criteria for many clinical trails include diagnosis history. It is challenging to accurately identify certain cohorts which meet specific diagnostic criteria for translational research without structured diagnosis codes for each individual animal. A small number of large veterinary centers hire dedicated coding staff to manually apply disease codes to clinical records, which is labor-intensive, while most veterinary clinics do not code the notes.[1] One potential solution that previous works have explored is to develop systems which automatically code veterinary notes. However, these approaches have been limited to specialized rule-based or machine learning-based models, which are tedious to design and do not easily generalize well to new formats of reports.

Large language models (LLMs) have the potential to serve as an effective method for veterinary information extraction. There is rising interest in studying large language models, commonly referred to as types of "foundation models" (models which can be adapted to many different tasks). LLMs have a large number of parameters and are typically pre-trained

on a large text corpus. They have shown promising performances on many NLP tasks, even in zero-shot and few-shot settings.[2,3] However, there is no study on how well those LLMs perform on analyzing veterinary notes. Besides, veterinary notes have shifted styles compared with general text available on Internet. For example, the vocabulary used is different and many acronyms are included. Given the pre-training corpus for most LLMs was sourced from Internet, veterinary notes are good examples to evaluate the performance of LLMs on atypical text.

Our contributions can be summarised as follows:

(1) We develop VetLLM to extract diagnostic information from veterinary notes and investigate its performance on dataset portions from two veterinary practices. Specifically we assess performance of models trained without any fine-tuning and with fine-tuning.
(2) We empirically show that LLMs can achieve promising performance on the task of diagnosis extraction from veterinary notes. Base LLMs without finetuning can achieve reasonable performances under zero-shot settings. For example, Alpaca-7B can achieve an zero-shot F1 of 0.538 when evaluated on CSU test data.
(3) Fine-tuned VetLLM perform better by a large margin compared with strong state-of-the-art supervised models. When evaluating on external test data, VetLLM outperforms the VetTag model by 21% and 8% in F1 score and exact match score respectively.
(4) We find finetuning LLM for the diagnosis extraction task is data-efficient. More specifically, using 200 notes can outperform supervised models trained with more than 100,000 notes in terms of F1 score.
(5) We detail a new paradigm for processing medical text in section 5.3, and the findings show the superiority of this new paradigm which leverages LLMs. Code will be available at `https://github.com/stanfordmlgroup/VetLLM`.

2. Related Work

Many previous studies have studied the automatic information extraction from clinical notes, including MetaMap,[4] statistical modeling,[5] text CNN,[6] and long-short-term memory network (LSTM).[7] More specifically, DeepTag[1] and VetTag[8] are some previous work on this dataset. DeepTag extended a bidirectional LSTM architecture with a hierarchical loss function and achieved better performances.[1] VetTag further leveraged transformers architecture[9] and conducted large-scale pre-training on veterinary text, leading to the current state-of-the-art performances on this dataset.[8]

There has been some recent studies showing many of those LLM are "generalist" in the sense that they can perform reasonably well on a large variety of tasks across domains.[2,3] In the medical domain, LLM have shown promising performances for many tasks, including information extraction,[10] medical Q&A,[3,11,12] generating USMLE-style questions[13] and radiology reports.[14] There are also some commentaries on the potential and regulation of LLM for medical use cases.[15–18]

3. Methods

The task is to extract diagnosis from veterinary notes which are in the free-text form. The extraction task can be formulated as a multi-class multi-label classification problem. Specifically, for each disease, the model should output whether there is positive mention in the veterinary clinical note.

The development pipeline included data cleaning, model selection, prompt design, resolver design, model finetuning, and system evaluation.

3.1. *Data*

The DeepTag[1] dataset was used for the project. It contains over 100K expert labeled veterinary notes from the Colorado State University (CSU) and a private practice clinic (PP). Both CSU portion and PP portion used here were previously used for VetTag, and it's noteworthy that VetTag was also pre-trained on another much larger dataset. In this project, we selected nine most prevalent diseases for analysis due to computational constraints. These nine diseases covered at least one diagnosis in around 90% cases in both CSU and PP portion, and they covered around 60% to 70% of all top-level disease labels. We removed incomplete reports which are shorter than 200 characters after manual review to ensure data quality.

The CSU portion contains 112,557 veterinary notes from the Colorado State University College of Veterinary Medicine and Biomedical Sciences. Each note was labeled with a set of SNOMED-CT codes by veterinarians at Colorado State. Colorado State is a tertiary referral center with an active and nationally recognized cancer center. We kept the same train/val/test as VetTag for fair comparison.

The PP portion contains 586 discharge summaries curated from a commercial veterinary practice located in Northern California, and six notes were removed due to incompleteness. Two veterinary experts applied SNOMED-CT codes to these records. Records with coding discrepancies were reviewed by both coders to reach a consensus on each record. This dataset is drastically different from the CSU dataset. PP notes are written often in an informal style, evidenced by their shorter length and usage of abbreviations. The PP data also has a different diagnosis distribution compared to a specialized academic cancer center CSU. It is of note that all notes in the PP portion are used for testing serving as an external validation dataset.

Table 1 shows the details of the dataset. Here is one example of veterinary note from PP portion together with the labels: *cried at home 8 body condition not drinking excess not urinating more frequently appetite is normal energy level is good skin is normal heart auscultates normal abnormal findings pain over l-s pain pulling hips back x rays ventral dorsal hips cauda equina looked perfect* **Expert annotated diseases:** 'Hypersensitivity condition', 'Propensity to adverse reactions'

Given both of these datasets are private and the data usage agreement prohibits data sharing with third parties, only models hosted locally can be used to analyze the data.

3.2. *Models*

Alpaca-7B and VetLLM Alpaca-7B[19] was used as the base LLM model as it has been instruction fine-tuned and is publicly available. Furthermore, a subset of CSU training split

Table 1. Descriptive statistics of the DeepTag dataset

	CSU	PP
# of notes	112,557	580
Size of test split	5483	580
Avg # of words	368	253

was used to further fine-tune Alpaca-7B using low-rank adaption,[20] leading to VetLLM. The details of fine-tuning was discussed in Section 3.4. The temperature for both Alpaca-7B and VetLLM was set to zero to allow reproducibility.

VetTag The supervised baseline model was the one developed in the VetTag paper, achieving state-of-the-art performances on the dataset.[8] It was pre-trained on a large corpus of unlabelled veterinary notes (917,665 notes) using casual language modeling, and then fine-tuned using the training split (101,301 notes) of CSU portion. The prediction logits from VetTag were obtained from the VetTag team, and the logits corresponding to the nine diseases were extracted to calculate the metrics.

KeywordMatch Another baseline model was to use keyword matching. The synonyms of the diseases were retrieved using WordNet, and fuzzy matching with the partial ratio metric was used. The model would return positive if the partial ratio between the veterinary note and the disease names was above 80%.

In short, four models would be compared: Alpaca-7B (LLM baseline), VetLLM (fine-tuned LLM), VetTag (supervised baseline) and KeywordMatch.

3.3. *Prompts and Resolvers*

The guiding principle for prompt design is to follow the format of the instruction tuning set and to be clear and specific. We just tried a small number of prompts on ten notes from the CSU validation split, and the main metric was whether the output was easily resolvable. Figure 1 shows the prompt used along with one example input and output from the VetLLM. The prompt queried the LLM with one disease each time rather than querying the LLM to list down all diagnosis. This design choice greatly simplified the resolver design. The query was conducted on two A4000 GPUs.

After getting the response from LLM, a resolver was utilized to convert the text response into a structured prediction. The resolver used in this study was simple, and it first converted the decoded text response into lower case and stripped any trailing space on the left. A positive prediction was rendered if the resultant string started with "yes", and a negative prediction was rendered if it started with "no". Otherwise, the case was rendered as un-resolvable.

3.4. *Finetuning*

A subset of 5,000 notes were randomly sampled from the CSU training split, and this subset was used for fine-tuning Alpaca-7B using low-rank adaption (LoRA)[20] and four A4000 GPUs. We chose LoRA as it generally provides superior performances and induces no extra inference overhead. The fine-tuning samples were generated using the same prompt template described

Prompt Template

Below is an instruction that describes a task, paired with an input that provides further context.
Write a response that appropriately completes the request.

Instruction:
Answer the following yes/no question based on the veterinary note delimited by triple backticks:

Input:
```{text}```

Does this animal have {disease}?

Response:

Example Input to VetLLM

Below is an instruction that describes a task, paired with an input that provides further context.
Write a response that appropriately completes the request.

Instruction:
Answer the following yes/no question based on the veterinary note delimited by triple backticks:

Input:
```tt w=12.9 lbs itching constantly flea allergy dermatitis 7 body condition not drinking excess not urinating more frequently appetite is normal energy level is good skin is normal heart auscultate s normal abnormal findings lots of hair loss fleas found other cat treatment , prescription dex dro ps disc flea allergy dermatitis and not spraying yard roos lmor 8 / 28```

Does this animal have Hypersensitivity condition?

Response:

Example Output from VetLLM

yes</s>

Fig. 1. Prompt Template with Example Input and Output from VetLLM

in the previous section. The hyper-parameters were included in Appendix A. A subset of 200 notes were randomly sampled from the CSU validation split to form the validation set for fine-tuning. An early stopping callback with a patience of five was added.

To study the data efficiency of fine-tuning, the 5,000 notes subset was further sampled into 2000, 1000, 500 and 200 notes sequentially. Consequently, each subset of a smaller size is strictly a subset of the one of a larger size. And these subsets were each used as the fine-tuning set. In short, five fine-tuned Alpaca-7B models were trained.

3.5. *Evaluation Metrics*

As a multi-class multi-label classification problem, there were metrics for both the overall prediction and each individual class. More specifically, each model was evaluated based on exact match (EM, the fraction of notes where the algorithm's predicted diagnoses exactly

Table 2. Quantitative evaluation on classification

Model	CSU				PP			
	Exact Match	Precision	Recall	F1	Exact Match	Precision	Recall	F1
VetLLM	$53.5\% \pm 0.7\%$	0.726	0.774	0.747 ± 0.004	$38.0\% \pm 2.1\%$	0.661	0.630	0.637 ± 0.015
Alpaca-7B (zero shot)	$34.0\% \pm 0.6\%$	0.604	0.527	0.538 ± 0.005	$22.0\% \pm 1.7\%$	0.485	0.375	0.389 ± 0.017
VetTag (supervised)	49.3% ±0.7%	0.798	0.492	0.592 ±0.006	30.1% ±1.9%	0.680	0.344	0.422 ±0.018
KeywordMatch	29.1%	0.442	0.002	0.003	24.9%	0.050	0.006	0.010

match the expert diagnoses), precision (the fraction of notes with positive predictions that match the expert diagnoses), recall (the fraction of notes where the expert diagnoses are successfully retrieved), and F1 (the harmonic mean of precision and recall). The last three metrics was macro-averaged across classes to get the overall metrics. The standard deviations of those metrics were calculated using bootstrapping with 1,000 re-samples.

4. Results

4.1. *Overall Evaluation on Classification*

Table 2 shows the quantitative evaluation results averaged across classes. When evaluating on the CSU portion, Alpaca-7B performs reasonably in a zero-shot manner, with only 6% gap in F1 compared with the supervised baseline. With VetLLM which was fine-tuned using 5,000 notes, the performances greatly improve, leading to a 21% boost in F1 and 19% boost in exact match score.

4.2. *Stratified Evaluation on Classification*

Figure 2 and 3 show the F1 metrics of three models evaluated on each class. They show the VetLLM model, fine-tuned from Alpaca-7B, outperforms the supervised VetTag model in each single class on both in-distribution data (CSU portion) and out-of-distribution data (PP portion). They also show significant improvements in performances in most classes after fine-tuning, and there is no degradation in any class after fine-tuning.

4.3. *Data-efficiency of Fine-tuning*

Figure 4 shows how the performance improves as the number of fine-tuning samples increase. It shows only using fewer than 200 notes can exceed performances of the supervised model, demonstrating the data-efficiency of fine-tuning LLM. It is of note the X-axis represents the number of veterinary notes used, so the size of fine-tuning set is nine times that the number of notes.

5. Discussion

The results show the promise of VetLLM for diagnosis extraction task from veterinary notes, which is inspiring. More broadly, they demonstrate the great potential of leveraging LLMs for processing medical text which is detailed in Section 5.3.

Although the performances are promising, they are sensitive to prompt design such as problem formulation, order of information and presence of extra information. In some cases,

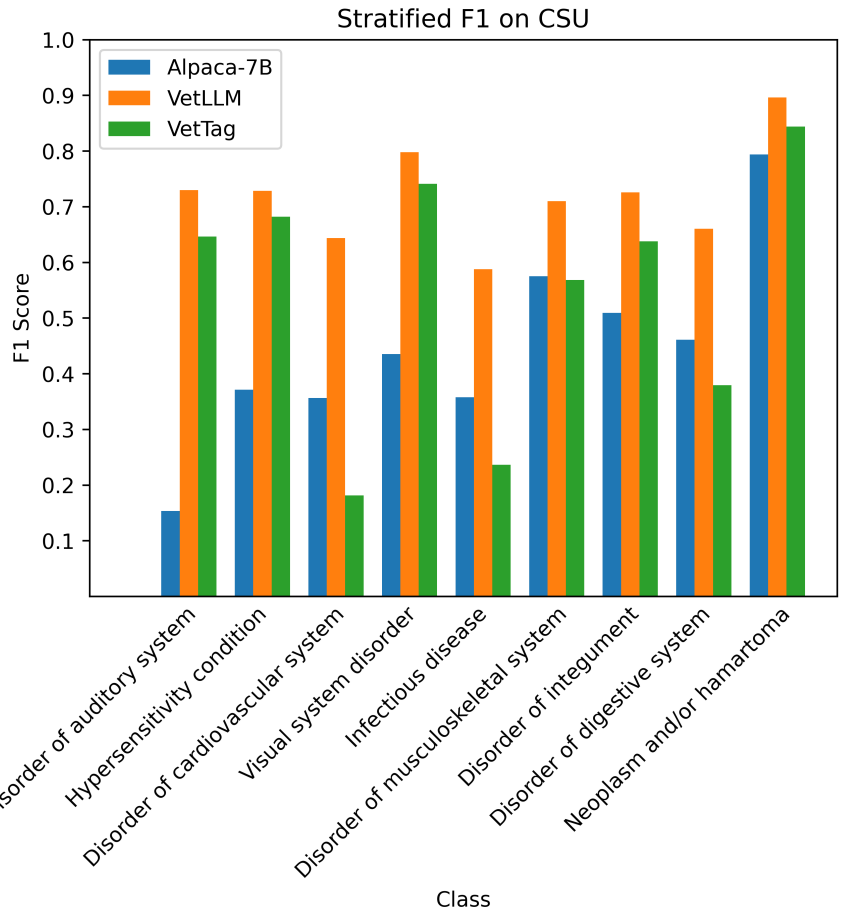

Fig. 2. Stratified F1 on CSU test data. Alpaca-7B is the base LLM model, and VetLLM is Alpaca-7B fine-tuned with 5,000 notes. VetTag is the state-of-the-art supervised model.

adding trailing spaces at the beginning of each line also affects the performances. It seems there is still no well-established systematic way of assessing LLM's sensitivity towards prompt designs, but the development of LLM is likely to benefit from ongoing research on AI alignment. Therefore, more comprehensive evaluation must be conducted or some post-hoc quality control measures must be taken if this system is to be deployed.

Also, the evaluation in this paper is limited to datasets from two centers in the United States. Veterinary notes from other veterinary medicine centers are likely to have different distributions which might affect performances.

5.1. *Error Analysis*

To gauge the knowledge embedded in LLM, the Alpaca-7B model was prompted to explain the top fourteen diseases. The responses from Alpaca-7B were included in Appendix B and manually reviewed in terms of relevance and factuality. The results indicate Alpaca can provide highly relevant and factually correct descriptions of diseases, hinting that the pre-training corpus might contain high-quality medical text describing various diseases.

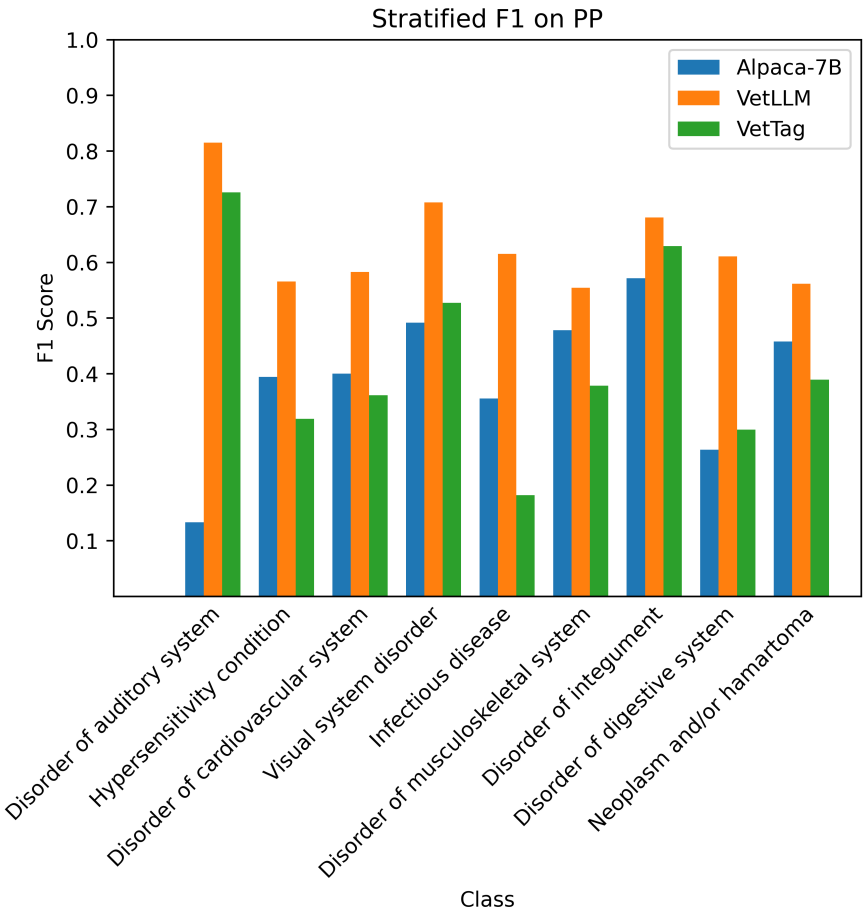

Fig. 3. Stratified F1 on PP test data. Alpaca-7B is the base LLM model, and VetLLM is Alpaca-7B fine-tuned with 5,000 notes. VetTag is the state-of-the-art supervised model.

Furthermore, the correlation between note length and performances were analyzed using the two portions, and the results are shown in Figure 5 and 6. The note lengths were binned into five quantiles. Based on the results, the exact match score is negatively correlated with the note length, while the trends for F1 score seem inconsistent.

5.2. *Computational Costs*

In the era of large models, computational costs and environmental impact of model training and inference have become more concerning. All estimates in this section are in the settings of four NVIDIA RTX A4000 GPUs launched in April 2021, and each A4000 has 16GB GPU memory. VetLLM was fine-tuned from Alpaca-7B using 5,000 notes, and the fine-tuning took around 48 hours with a micro batch size of one.

One limitation of VetLLM is it requires multiple pass for multi-class classification, while traditional supervised models can generate multi-class predictions with a single pass by using multiple neurons in the last layer. A single inference pass for VetLLM takes around 0.3 seconds, and the model loading before first use takes around 15 seconds. It means VetLLM is likely to

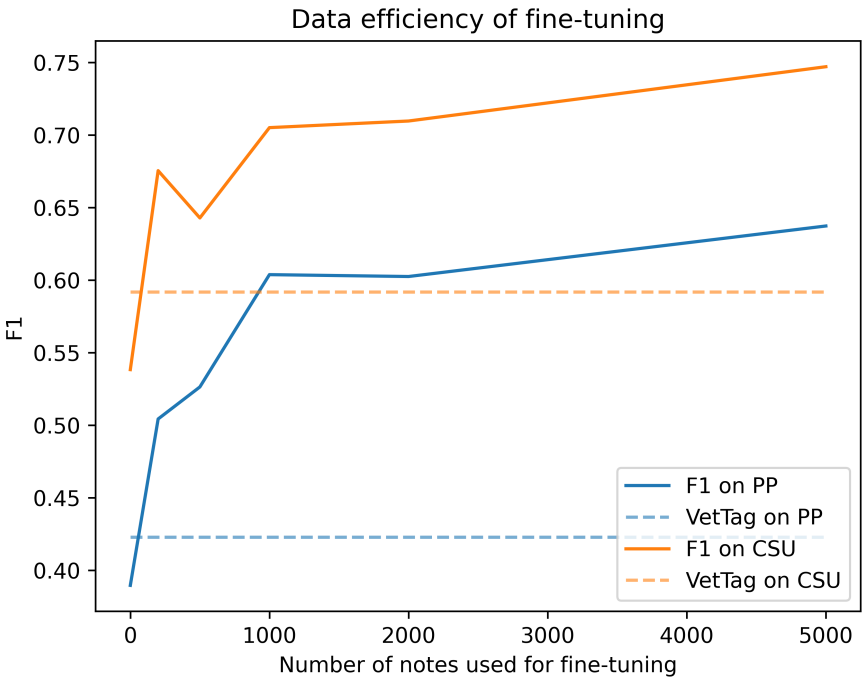

Fig. 4. Data efficiency plot. The number of notes here refer to the number of notes in the CSU training split used for fine-tuning Alpaca-7B. PP portion was only used for test. It is of note that VetTag used over 100,000 notes for fine-tuning.

have slower inference speed compared with traditional supervised models. Given the significant boost in performances and the application does not have strong real-time requirement, we think the increased inference time is reasonable. One mitigation, which we leave as future work, is to utilize a multi-label approach such as asking multiple questions in a single turn or asking the model to select all diseases present in the veterinary note.

5.3. *New Paradigm for Processing Medical Text*

In this paper, we demonstrate the potential of a new paradigm for processing medical text: starting with pre-trained large language models (LLM), then designing a prompt and resolver. The resolver interprets the output from the LLM and transforms the raw output into structured answers. After designing the prompt and resolver, the next step is to conduct a quick evaluation in a zero-shot or few-shot setting. If the performance is satisfactory, it is a good idea to proceed with more comprehensive evaluation and iteratively improve the prompt and resolver. If the performance is poor, it might be worth curating a small fine-tuning dataset and utilizing data-efficient techniques like LoRA to fine-tune the LLM. Evaluation and iterative refinement can be conducted after the fine-tuning.

Thanks to the great contributions from various communities, most of these steps have been implemented in various library packages or are available as API calls, thereby speeding up the entire pipeline. The traditional pipeline for processing medical text involves curating a large training dataset and training a specialized model via supervised learning. This pipeline

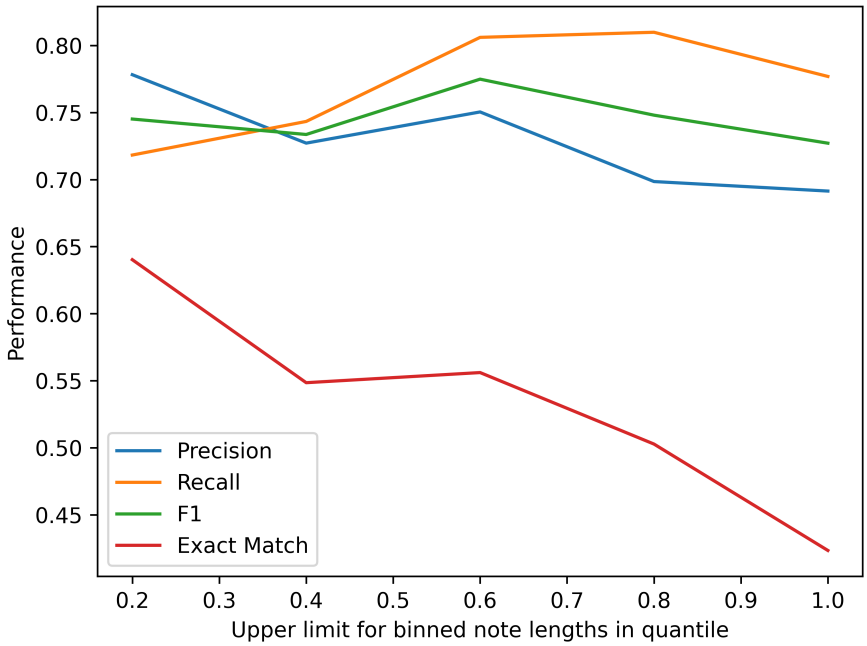

Fig. 5. Performances of VetLLM in terms of note length on CSU test data. The X-axis refers to the upper quantile of note length for each binned group. Exact match is a baseline evaluation metric, and F1 is used more frequently in practice.

tends to require significant resources, with the process often spanning several months or even years. This new paradigm might lower the barriers to building some interesting applications, with many potentially developed within weeks.

Beyond the great performance and fast iteration, another advantage of this new paradigm is the ability to easily expand classification categories. For example, the prompt can be modified to extract diagnosis of other diseases. In contrast, traditional supervised models might require extensive fine-tuning to include new classes.

6. Conclusion

In this study, large language models were used for diagnosis extraction task from veterinary notes. With fine-tuning only on a small number of notes, VetLLM outperform strong supervised models significantly. Given the time constraints, simple prompts and resolvers were used in the study. Richer prompt strategies can be explored, and robustness towards prompt variations should be examined.

In a broader sense, this project has shown the potential of LLMs to work on clinical data and be efficiently fine-tuned to achieve strong performances on downstream tasks. Although this study is limited to veterinary notes, we believe the new paradigm detailed in section 5.3 is generally applicable. Therefore, it is interesting to evaluate the performances of base LLMs and fine-tuned LLMs in other medical applications including human clinical notes. Furthermore, it might be interesting to conduct similar assessment using more advanced models or more domain-specific ones.

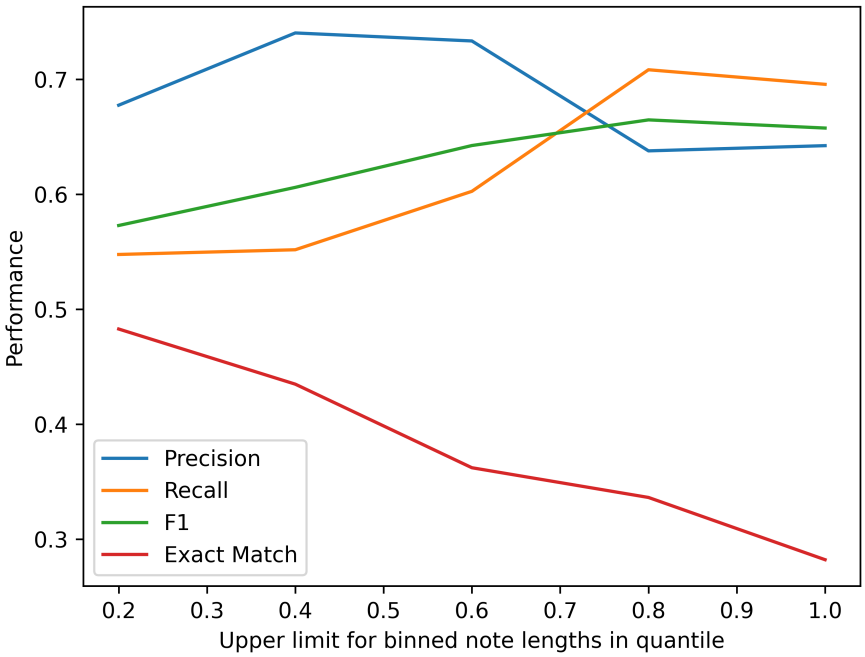

Fig. 6. Performances of VetLLM in terms of note length on PP test data. The X-axis refers to the upper quantile of note length for each binned group. Exact match is a baseline evaluation metric, and F1 is used more frequently in practice.

Acknowledgments

Yixing Jiang is supported by National Science Scholarship (PhD). We want to thank Anh (Tom) Nguyen and Tulika Jha for helping with some experiments. Besides, we want to thank Yuhui Zhang, Allen Nie, Federico Bianchi and Bryan He for helping with dataset and model access.

Appendix A. Hyper-parameters used for fine-tuning

Table A1 shows the hyper-parameters used for LoRA fine-tuning.

Table A1. Hyper-parameters used for fine–tuning

Name	Value
batch_size	32
max_epochs	3
cutoff_len	2048
train_on_inputs	False
lora_r	8
lora_alpha	16
lora_dropout	0.05
lora_target_modules	[”q_proj”, ”v_proj”]

Appendix B. Alpaca-7B' responses to disease definition questions

- Disorder of auditory system : Disorder of the auditory system is a broad term that refers to any disorder of the ear or hearing system that results in a hearing impairment or loss. This can include conditions such as conductive hearing loss, sensorineural hearing loss, mixed hearing loss, and other disorders that affect the ability to hear.
- Hypersensitivity condition : Hypersensitivity is an abnormal increase in the body's reactivity to a substance or stimulus, such as an allergy or an autoimmune disorder. Hypersensitivity reactions can range from mild to severe and can affect the skin, respiratory tract, gastrointestinal tract, and other organs. Symptoms may include rash, hives, eczema, congestion, coughing, wheezing, and other respiratory symptoms, as well as digestive issues.
- Disorder of cardiovascular system : Disorder of the cardiovascular system is any abnormality of the heart, blood vessels, or blood that affects how the body circulates blood. This can include conditions such as high blood pressure, heart disease, stroke, and blood clots.
- Visual system disorder : Visual system disorder is a condition that affects the way the eyes send and process visual information. It can cause problems with vision, such as blurred vision, double vision, or difficulty with eye coordination and tracking. It can also lead to headaches, eyestrain, and other vision-related problems.
- Infectious disease : Infectious disease is a broad term used to describe any disease that is caused by the presence, growth, and/or activity of a pathogen, such as a virus, bacteria, parasite, or fungus. These pathogens can cause illnesses ranging from mild to severe, and can even be fatal. Infectious diseases are spread through contact with infected individuals, contact with contaminated surfaces, or contact with contaminated food or water.
- Disorder of musculoskeletal system : Disorder of the musculoskeletal system is a general term that describes any injury or disease that affects the muscles, bones, joints, and associated structures. Examples of disorders of the musculoskeletal system include fractures, sprains, strains, tendinitis, bursitis, and muscle pain.
- Disorder of integument : Disorder of the integument is a general term used to describe a group of skin disorders that affect the structure or function of the skin. These disorders can range from mild forms such as dry skin or eczema, to more severe forms such as psoriasis or dermatitis.
- Disorder of digestive system : Disorder of the digestive system is any problem with the digestive tract that impairs its functioning and can cause symptoms such as abdominal pain, nausea, vomiting, diarrhea, constipation, and other complications. Common disorders of the digestive system include gastroenteritis, irritable bowel syndrome, inflammatory bowel disease, and cancer.
- Neoplasm and/or hamartoma : Neoplasm is a general term for a tumor or growth that is caused by abnormal cellular proliferation, while hamartoma is a benign tumor composed of normal tissue.

References

1. A. Nie, A. Zehnder, R. L. Page, Y. Zhang, A. L. Pineda, M. A. Rivas, C. D. Bustamante and J. Zou, Deeptag: inferring diagnoses from veterinary clinical notes, *NPJ digital medicine* **1**, p. 60 (2018).
2. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, Language models are few-shot learners, *Advances in neural information processing systems* **33**, 1877 (2020).
3. H. Nori, N. King, S. M. McKinney, D. Carignan and E. Horvitz, Capabilities of gpt-4 on medical challenge problems, *arXiv preprint arXiv:2303.13375* (2023).
4. A. R. Aronson and F.-M. Lang, An overview of metamap: historical perspective and recent advances, *Journal of the American Medical Informatics Association* **17**, 229 (2010).
5. M. Subotin and A. R. Davis, A method for modeling co-occurrence propensity of clinical codes with application to icd-10-pcs auto-coding, *Journal of the American Medical Informatics Association* **23**, 866 (2016).
6. Y. Zhang and B. Wallace, A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification, *arXiv preprint arXiv:1510.03820* (2015).
7. Z. C. Lipton, D. C. Kale, C. Elkan and R. Wetzel, Learning to diagnose with lstm recurrent neural networks, *arXiv preprint arXiv:1511.03677* (2015).
8. Y. Zhang, A. Nie, A. Zehnder, R. L. Page and J. Zou, Vettag: improving automated veterinary diagnosis coding via large-scale language modeling, *NPJ digital medicine* **2**, p. 35 (2019).
9. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* **30** (2017).
10. M. Agrawal, S. Hegselmann, H. Lang, Y. Kim and D. Sontag, Large language models are few-shot clinical information extractors, in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.
11. K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl *et al.*, Large language models encode clinical knowledge, *arXiv preprint arXiv:2212.13138* (2022).
12. T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo *et al.*, Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models, *PLoS digital health* **2**, p. e0000198 (2023).
13. S. L. Fleming, K. Morse, A. M. Kumar, C.-C. Chiang, B. Patel, E. P. Brunskill and N. Shah, Assessing the potential of usmle-like exam questions generated by gpt-4, *medRxiv* , 2023 (2023).
14. D. Van Veen, C. Van Uden, M. Attias, A. Pareek, C. Bluethgen, M. Polacin, W. Chiu, J.-B. Delbrouck, J. M. Z. Chaves, C. P. Langlotz *et al.*, Radadapt: Radiology report summarization via lightweight domain adaptation of large language models, *arXiv preprint arXiv:2305.01146* (2023).
15. R. Li, A. Kumar and J. H. Chen, How chatbots and large language model artificial intelligence systems will reshape modern medicine: Fountain of creativity or pandora's box?, *JAMA Internal Medicine* (2023).
16. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan and D. S. W. Ting, Large language models in medicine, *Nature Medicine* , 1 (2023).
17. B. Meskó and E. J. Topol, The imperative for regulatory oversight of large language models (or generative ai) in healthcare, *NPJ Digital Medicine* **6**, p. 120 (2023).
18. S. Gilbert, H. Harvey, T. Melvin, E. Vollebregt and P. Wicks, Large language model ai chatbots require approval as medical devices, *Nature Medicine* , 1 (2023).
19. R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang and T. B.

Hashimoto, Stanford alpaca: An instruction-following llama model `https://github.com/tatsu-lab/stanford_alpaca`, (2023).
20. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, Lora: Low-rank adaptation of large language models, *arXiv preprint arXiv:2106.09685* (2021).