Is That the Right Dose? Investigating Generative Language Model Performance on Veterinary Prescription Text Analysis

Brian Hur[1] **Lucy Lu Wang**[1,2] **Laura Hardefeldt**[3] **Meliha Yetsigen**[1]
[1]University of Washington [2]Allen Institute for AI [3]University of Melbourne
{hurb, lucylw, melihay}@uw.edu, laura.hardefeldt@unimelb.edu.au

Abstract

Optimizing antibiotic dosing recommendations is a vital aspect of antimicrobial stewardship (AMS) programs aimed at combating antimicrobial resistance (AMR), a significant public health concern, where inappropriate dosing contributes to the selection of AMR pathogens. A key challenge is the extraction of dosing information, which is embedded in free-text clinical records and necessitates numerical transformations. This paper assesses the utility of Large Language Models (LLMs) in extracting essential prescription attributes such as dose, duration, active ingredient, and indication. We evaluate methods to optimize LLMs on this task against a baseline BERT-based ensemble model. Our findings reveal that LLMs can achieve exceptional accuracy by combining probabilistic predictions with deterministic calculations, enforced through functional prompting, to ensure data types and execute necessary arithmetic. This research demonstrates new prospects for automating aspects of AMS when no training data is available.

Clinical Note Inputs

Consultation Note:
History: Realised on Saturday that there was a wound on left fore leg...
Examination: Not limping in consult room...1cm wound that has been filled with granulation tissue. Swelling approx 0.5cm
Assessment: Wound over left fore limb...
Plan: Recheck in a week.

Prescription Label Information:
Item Label: Dog 14.00 x Amoxyclav Tabs 250Mg Give half tablet twice a day
Item Name: Amoxyclav Tabs 250Mg (100)
Weight: 10kgs
Units Dispensed: 14.0

Inferred Labels for Evaluation

Ingredient: Amoxicillin Clavulanate
Indication: Traumatic Injury
Frequency: 2 (daily doses)
Medication Size: 250 (mgs)
Dose Unit Size: 0.5 (tablets per dose)
Duration: 14 (days)
Dose: 12.5 (mg/kg)

Figure 1: Example of consultation and prescription note along with inferred labels.

1 Introduction

Antimicrobial resistance (AMR) has become a major public health concern, as antimicrobials are steadily losing their effectiveness in combating bacterial infections (O'Neill, 2016). AMR is not limited to human medicine; it is also a growing issue among animals (Ekakoro et al., 2022; Cummings et al., 2015), who can acquire and transmit multidrug-resistant pathogens to humans (Guardabassi et al., 2004). Antimicrobial Stewardship (AMS), which has demonstrated effectiveness in improving antimicrobial use in both human and animal healthcare (Davey et al., 2017; Hardefeldt et al., 2022), aims to optimize antimicrobial use to curtail the development and spread of AMR. Accurate dosing is part of this strategy, as overdosing can lead to toxicity and under-dosing can be particularly perilous as it can select for AMR organisms and lead to poor therapeutic outcomes (Roe et al., 2012; Grill and Maganti, 2011). A pragmatic way to improve dosing accuracy and optimizing antimicrobial use is through decision support systems in clinical settings (Hardefeldt et al., 2018b,a), where targeted dosing recommendations can be made in real-time.

Recent developments in Large Language Models (LLMs) introduce compelling opportunities for automated information extraction and decision support (Bubeck et al., 2023; Nori et al., 2023), as these models obviate the need for extensive labeled data (Brown et al., 2020). Such models can potentially furnish clinicians with data-driven counsel on optimal antimicrobial selection, treatment du-

ration, and dosing intervals, tasks that have been historically reliant on extensive labor-intensive labeled data compilation (Uzuner et al., 2010; Tao et al., 2017). To realize the potential for LLMs for extracting prescription elements, an essential step is empirical assessment of their ability to accurately extract relevant information from clinical text. Given the idiosyncratic nature of LLM training, which leverages instruction tuning rather than conventional training paradigms, it becomes vital to also scrutinize configuration variances for performance optimization (Zheng et al., 2023). Additionally, while the task of extracting elements out of prescriptions was explored in shared tasks such as the 2010 i2b2 challenge (Uzuner et al., 2011), these studies only evaluate the ability to extract text spans without performing numerical conversion. Converting text spans to numerical representations and performing necessary calculations to understand the dose and duration of a given medication are also essential to optimize antimicrobial use. Studies performing such numerical conversions rely on rules-based methods which are notoriously brittle, and only one study we identified made the corresponding algorithm available (Karystianis et al., 2016).

We leverage the VetCompass Australia (McGreevy et al., 2017) corpus, which comprises over 50 million clinical notes from over 200 veterinary clinics across Australia, as our primary data source. Our goal is to extract key information such as the active ingredient, the indication for antimicrobial use, and the dose and duration of the therapy. We assess the performance of LLMs in zero-shot and few-shot learning scenarios for extracting this critical information. By exploring the feasibility of applying LLMs to the VetCompass dataset, we seek to understand their potential in aiding dosing recommendations to support AMS. Specifically:

- We construct a veterinarian-labeled evaluation dataset of 200 clinical notes to study medication dosage extraction from veterinary notes;
- Using silver labels generated by a baseline BERT-based ensemble model to provide training examples, we benchmark the performance of LLMs against the baseline model for extracting medication dosage information, their indications, and active ingredients;
- While we demonstrate LLMs' proficiency in element extraction for dose and duration calculations, they falter at arithmetic operations crucial for deriving these elements (Yuan et al., 2023). We introduce methods to overcome this using functional prompting to combine the probabilistic predictions from the LLMs with deterministic calculations for labelling dosing elements in zero- or few-shot settings.[1]

2 Task & Dataset

We investigate the task of dose information extraction from veterinary clinical notes. Given textual clinical notes, the task is to extract seven labels, including five entity labels: active ingredient, clinical indication, frequency, medication size, and dosage unit size, along with two derived labels: dose and duration. A sample clinical note, prescription label, and the inferred target labels are illustrated in Figure 1. Prescription label information is provided as an input for all extractions except indication, for reasons of document length. For indication, the model is also given the set of potential indications; for ingredient, the set of potential ingredients. We evaluate accuracy based on exact match between the output label and the ground truth label.

Data Extraction and Label Creation We assemble a subset of 1500 clinical records sourced from VetCompass Australia (McGreevy et al., 2017), focusing on cases where patients received oral antimicrobial treatments as outlined in Hur et al. (2019). To facilitate further calculations, the patient's weight in kilograms and the total quantity of medication dispensed are also extracted from structured textual fields within the clinical records (Appendix Table 3). We extract the inferred label elements shown in Figure 1 using `RxVetBERT`, an ensemble model introduced in prior work (Hur et al., 2020, 2022); these inferred labels are used as silver labels for in-context learning examples. The extracted records and labels are partitioned into 1000 records for training, 300 for development, and 200 for test. The test set is reserved exclusively for the final evaluation stage after all prompts have been refined and optimized.

Gold Test Set Annotations To ensure label accuracy in the test set, two expert veterinarians manually annotated the data. Inter-annotator agreement was evaluated using exact match F1 scores. Initial IAA F1 was 0.8 for Indication, 0.985 for Dose, and

[1]The code and select models used in this study available at https://github.com/havocy28/prescription-text-analyzer

1.0 for Duration, Ingredient, Medication Unit Size and Dose Unit Size. High agreement in dosage and ingredient categories was due to their objectivity. Consensus on indication is more challenging, particularly when multiple clinical events complicated interpretation—e.g., in scenarios involving post-operative complications following traumatic injury, the indication could be correctly interpreted as either the initial injury or subsequent complications. Any annotation discrepancies were resolved through consensus discussion.

Indication and Ingredient Labels Indication labels are based on a subset of Veterinary Nomenclature (VeNOM) codes, a specialized adaptation of SNOMED for veterinary medicine (Brodbelt, 2019). We use the subset of 52 curated by (O'Neill et al., 2019), of which 23 appear in our test set. Ingredient labels are based on unique antimicrobial agents from VetCompass, which consist of 49 unique ingredients, 9 of which occur in our test set.

Dosing Elements Frequency indicates the amount of times per day a dose is given. It must be a numerical value such that it can be used in dosing calculations (e.g., 'twice daily' is converted to 2). Dose unit size indicates the amount of medication and must also be converted into a numerical value (e.g., 'half of a tablet' is converted to 0.5).

The medication dose and duration can then be calculated using the formulae:

$$\text{Dose} = \frac{D \times M}{W} \quad \text{Duration} = \frac{T}{F \times D}$$

where D is the Dose Unit Size (number of tablets or volume of liquid), M the Medication Size (tablet size [in mg]), W the Weight of Patient (in kg), T the Total Units Dispensed, and F the Administration Frequency. The dose calculation is designed to tailor the medication dose to the individual's mass to achieve the optimal therapeutic efficacy while minimizing the risk of toxicity. This is particularly important for veterinary and pediatric patients, where the difference in mass between patients can vary greatly (Waldman et al., 2008).

3 Methodology

We benchmark three LLMs on this task: `GPT-3.5` (Brown et al., 2020), `GPT-4` (OpenAI, 2023a), and `LLAMA2-70B` (Touvron et al., 2023), against the baseline ensemble model (`RxVetBERT`), which combines rule-based methods and `VetBERT` as described in previous works (Hur et al., 2020, 2022).

Prompt Settings We compare the following:

Zero-shot: Utilizes text from the clinical note and/or prescription label, along with the item name, weight of the patient, and the number of units dispensed as input. A prompt for the element being classified is included. No examples are provided.

Few-shot Random Examples: Incorporates randomly-sampled example prescriptions or examination text and inferred labels from the training set as in-context examples (Brown et al., 2020). To manage token limits, we include three labeled examples for prescription prompts and two for indication prompts—the examination text required for the indication prompt were much longer.

Few-shot Similar Examples: Instead of random examples, we use text similarity as a selection criterion to retrieve examples for in-context learning (Zhang et al., 2023; Shi et al., 2023; Lewis et al., 2021). For the retriever, we employ a distilled `SBERT` model (Wang et al., 2020) to encode text and retrieve examples based on cosine similarity.

Functional Prompting: In the zero-shot setting, we leverage functional prompting (OpenAI, 2023b) with `GPT-3.5` and `GPT-4` to combine probabilistic outputs with rule-based calculations, enforcing data types for extracted prescription attributes and executing formulaic calculations for Dose and Duration, as detailed in §2. We compare these results with `LLAMA2-70B`'s configurations for extracting dose unit size and frequency. Additionally, we finetune a `VetBERT` model (Hur et al., 2020) with silver label data to isolate these attributes, and perform deterministic calculations for dose and duration.

Prompt Tuning To improve the performance of calculating the dose and duration of therapy, we include the formulas for the dose and duration calculations as part of the prompt. The prompts used for evaluation were additionally optimized using the framework proposed by Yang et al. (2023) to iteratively generate a set of prompts using `GPT-4`, test those prompts on a subset of records from the training set until no improvements were observed after multiple iterations, and keep the prompt with highest accuracy for each element. Final prompts can be found in Appendix A.2.

Postprocessing We remove non-numerical text and retain the first float in the model's output for enhanced accuracy, except for indication and ingredient which are expected to be strings.

	Ingredient	Indication	Dose	Duration	Frequency	Dose Unit Size
RxVetBERT	100	80.0	89.1	88.0	97.0	89.0
Few-Shot Similar Examples						
GPT-3.5	97.5	56.5	29.5	70.0	98.0	98.0
GPT-4	99.5	75.0	85.0	91.0	98.5	99.5
LLAMA2-70B	94.0	9.0	12.5	58.0	97.5	95.5
Few-Shot Random Examples						
GPT-3.5	67.0	73.5	26.0	61.0	98.5	97.0
GPT-4	100	73.5	88.5	84.5	98.5	100
LLAMA2-70B	42.0	27.5	9.5	61.0	97.5	92.5
Zero-Shot						
GPT-3.5	80.5	35.0	3.5	52.5	12.0	21.0
GPT-4	97.5	69.5	24.0	75.5	97.5	55.0
LLAMA2-70B	21.0	0.0	5.0	57.5	98.0	59.5

Table 1: Model accuracy (%) across multiple settings, benchmarked against RxVetBERT.

	Dose	Duration	Freq.	Dose Unit Size
Finetuned				
VetBERT	90.0	88.0	97.0	90.5
Few-Shot Similar Examples				
LLAMA2-70B	95.5	93.5	97.5	95.5
Zero-Shot				
GPT-3.5	94.5	92.5	98.0	98.0
GPT-4	99.5	98.0	98.5	99.5

Table 2: Evaluation of GPT-3.5 and GPT-4 in a zero-shot setting using functional prompts to enforce numerical data types, compared to VetBERT trained on silver labels and LLAMA2-70B in the Few-Shot Similar setting. Dose and Duration are computed deterministically for all model variants.

4 Results and Discussion

Overall, our experiments find that LLMs are highly effective at extracting and interpreting numerical elements (e.g., Frequency and Dose Unit Size) necessary to calculate the dose and duration (Table 1). GPT-3.5 and GPT-4 show high accuracy in the zero-shot setting and LLAMA2-70B using in-context learning examples. For all models, integrating probabilistic outputs with deterministic calculations through functional prompting achieves much higher accuracy for dose and duration values compared to directly prompting models for these values (Table 2). Functional prompting provides an effective way to ensure more reliable outcomes in tasks requiring numerical computations when those computations can be explicitly described.

The much smaller finetuned VetBERT achieves modest results, similar to silver label accuracy for dose unit size and frequency of administration, which suggests it could achieve higher accuracy with more accurate training labels. For active ingredient, we find LLMs to be highly effective with appropriate prompt tuning. Error analysis of initial results found that there were many errors for multi-ingredient medications, e.g., Amoxycillin Clavulanate incorrectly identified as Amoxycillin. Through prompt tuning, we identified the most effective way to overcome this as including the following prompt text: "focus on the active ingredients and note them all if they were present."

Nonetheless, accurately pinpointing the primary indication for antimicrobial administration continues to be a challenging task, as evidenced by the lower inter-annotator agreement score for indications as discussed in §2, and the relatively poor performance of LLMs on this subtask. A closer examination of the errors reveals that they largely occurred in instances where the indication is ambiguous, similar to the complications noted earlier. Refining the labeling schema for indications is a promising avenue for mitigating this issue.

Conclusion This paper provides a framework for LLMs to extract essential prescription data from veterinary text, such as dose, duration, and active ingredients for supporting AMS efforts. We overcome limitations in calculating elements such as dose and duration by integrating probabilistic outputs with deterministic calculations through functional prompting, even in zero-shot settings. Future work should consider evaluation for human clinical applications, given the potential contributions of this approach to broader healthcare.

Limitations

The efficacy of in-context learning for models in the few-shot similar setting may be constrained by the precision of `RxVetBERT`, which was employed to furnish the examples used in the prompts. Random sampling was used to create a test set mirroring the full dataset population; this limited the diversity of specific disease syndromes in the test data, and may not provide a complete assessment of the models' capabilities.

LLMs are still prone to errors, even though they demonstrate high performance on evaluation datasets. When using an LLM for clinical decisions, it is critical that the final decisions involve clinicians as LLMs in their current state may still fall short. While our framework excels in identifying active ingredients, it faces challenges in ascertaining exact indications for medication, a more subjective task, signaling a potential direction for future work.

Acknowledgements

This work was supported in part by the National Institutes of Health, National Library of Medicine (NLM) University of Washington Biomedical Informatics and Data Science Research Training Program (Grant Nr LM 007442). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was undertaken with the assistance of information and other resources from the VetCompass Australia consortium under the project "VetCompass Australia: Big Data and Real-time Surveillance for Veterinary Science", which is supported by the Australian Government through the Australian Research Council LIEF scheme (LE160100026).

References

David Brodbelt. 2019. VeNom Coding – VeNom Coding Group.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Kevin J. Cummings, Victor A. Aprea, and Craig Altier. 2015. Antimicrobial resistance trends among canine Escherichia coli isolates obtained from clinical samples in the northeastern USA, 2004-2011. *The Canadian Veterinary Journal = La Revue Veterinaire Canadienne*, 56(4):393–398.

Peter Davey, Charis A Marwick, Claire L Scott, Esmita Charani, Kirsty McNeil, Erwin Brown, Ian M Gould, Craig R Ramsay, and Susan Michie. 2017. Interventions to improve antibiotic prescribing practices for hospital inpatients. *The Cochrane Database of Systematic Reviews*, 2017(2):CD003543.

John E. Ekakoro, G. Kenitra Hendrix, Lynn F. Guptill, and Audrey Ruple. 2022. Antimicrobial susceptibility and risk factors for resistance among Escherichia coli isolated from canine specimens submitted to a diagnostic laboratory in Indiana, 2010-2019. *PloS One*, 17(8):e0263949.

Marie F Grill and Rama K Maganti. 2011. Neurotoxic effects associated with antibiotic use: management considerations. *British Journal of Clinical Pharmacology*, 72(3):381–393.

Luca Guardabassi, Stefan Schwarz, and David H. Lloyd. 2004. Pet animals as reservoirs of antimicrobial-resistant bacteriaReview. *Journal of Antimicrobial Chemotherapy*, 54(2):321–332.

L. Y. Hardefeldt, J. R. Gilkerson, H. Billman-Jacobe, M. A. Stevenson, K. Thursky, G. F. Browning, and K. E. Bailey. 2018a. Antimicrobial labelling in Australia: a threat to antimicrobial stewardship? *Australian Veterinary Journal*, 96(5):151–154. Https://doi.org/10.1111/avj.12677.

L. Y. Hardefeldt, B. Hur, S. Richards, R. Scarborough, G. F. Browning, H. Billman-Jacobe, J. R. Gilkerson, J. Ierardo, M. Awad, R. Chay, and K. E. Bailey. 2022. Antimicrobial stewardship in companion animal practice: an implementation trial in 135 general practice veterinary clinics. *JAC-Antimicrobial Resistance*, 4(1):dlac015.

Laura Y. Hardefeldt, J. R. Gilkerson, H. Billman-Jacobe, M. A. Stevenson, K. Thursky, K. E. Bailey, and G. F. Browning. 2018b. Barriers to and enablers of implementing antimicrobial stewardship programs in veterinary practices. *Journal of Veterinary Internal Medicine*, 32(3):1092–1099.

B. Hur, L. Y. Hardefeldt, K. Verspoor, T. Baldwin, and J. R. Gilkerson. 2019. Using natural language processing and VetCompass to understand antimicrobial usage patterns in Australia. *Australian Veterinary Journal*, 97(8):298–300.

Brian Hur, Timothy Baldwin, Karin Verspoor, Laura Hardefeldt, and James Gilkerson. 2020. Domain Adaptation and Instance Selection for Disease Syndrome Classification over Veterinary Clinical Notes. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 156–166, Online. Association for Computational Linguistics.

Brian Hur, Laura Y. Hardefeldt, Karin M. Verspoor, Timothy Baldwin, and James R. Gilkerson. 2022. Evaluating the dose, indication and agreement with guidelines of antimicrobial use in companion animal practice with natural language processing. *JAC-Antimicrobial Resistance*, 4(1):dlab194.

George Karystianis, Therese Sheppard, William G. Dixon, and Goran Nenadic. 2016. Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database. *BMC Medical Informatics and Decision Making*, 16.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv:2005.11401 [cs].

Paul McGreevy, Peter Thomson, Navneet K. Dhand, David Raubenheimer, Sophie Masters, Caroline S. Mansfield, Timothy Baldwin, Ricardo J. Soares Magalhaes, Jacquie Rand, Peter Hill, Anne Peaston, James Gilkerson, Martin Combs, Shane Raidal, Peter Irwin, Peter Irons, Richard Squires, David Brodbelt, and Jeremy Hammond. 2017. VetCompass Australia: A National Big Data Collection System for Veterinary Science. *Animals : an Open Access Journal from MDPI*, 7(10):74.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. ArXiv:2303.13375 [cs].

Jim O'Neill. 2016. Tackling drug-resistant infections globally: final report and recommendations. Report, Government of the United Kingdom.

OpenAI. 2023a. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

OpenAI. 2023b. OpenAI Platform - Function Calling.

Dan G. O'Neill, Alison M. Skipper, Jade Kadhim, David B. Church, Dave C. Brodbelt, and Rowena M. A. Packer. 2019. Disorders of Bulldogs under primary veterinary care in the UK in 2013. *PLOS ONE*, 14(6):e0217928.

Jada L. Roe, Joseph M. Fuentes, and Michael E. Mullins. 2012. Underdosing of common antibiotics for obese patients in the ED. *The American Journal of Emergency Medicine*, 30(7):1212–1214.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models. ArXiv:2301.12652 [cs].

Carson Tao, Michele Filannino, and Özlem Uzuner. 2017. Prescription extraction using CRFs and word embeddings. *Journal of Biomedical Informatics*, 72:60–66.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010. Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of the American Medical Informatics Association : JAMIA*, 17(5):519–523.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556.

Scott A. Waldman, Andre Terzic, Laurence J. Egan, Jean Luc Elghozi, Arshad Jahangir, Garvan C. Kane, Walter K. Kraft, Lionel D. Lewis, Jason D. Morrow, Leonid V. Zingman, Darrell R. Abernethy, Arthur J. Atkinson, Neal L. Benowitz, D. Craig Brater, Jean Gray, Peter K. Honig, Gregory L. Kearns, Barbara A. Levey, Stephen P. Spielberg, Richard Weinshilboum, and Raymond L. Woosley. 2008. *Pharmacology and Therapeutics: Principles to Practice*. Elsevier.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic

Compression of Pre-Trained Transformers. ArXiv:2002.10957 [cs].

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. 2023. Large Language Models as Optimizers. ArXiv:2309.03409 [cs].

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do Large Language Models perform in Arithmetic tasks? ArXiv:2304.02015 [cs].

Xuchao Zhang, Menglin Xia, Camille Couturier, Guoqing Zheng, Saravan Rajmohan, and Victor Ruhle. 2023. Hybrid Retrieval-Augmented Generation for Real-time Composition Assistance. ArXiv:2308.04215 [cs].

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. 2023. Progressive-Hint Prompting Improves Reasoning in Large Language Models. ArXiv:2304.09797 [cs].

A Appendix

A.1 Prompt Design Details

This section details the prompt template used and provides an example prompt. Each initial prompt was generated using `GPT-4` by prompting the model to generate 10 additional prompts for accomplishing the task. Each task was tested using `GPT-3.5` and the process was repeated until the additional prompts were no longer improving the performance after 3 successive iterations. For dose and duration, which also required the steps for performing arithmetic, the formula for the necessary arithmetic was provided.

The prompt templates were designed as follows:

```
{PROMPT}

{Examples} (omitted in zero-shot settings)

{Instance to label}
```

Here is an example prompt for dose under the few-shot random setting:

```
Output the dosage in mg/kg. Dose is determined
by multiplying the total dose units given per
administration multiplied by the size of the
medication in mg, dividing by the weight of the
patient in kg to determine the mg per kg

** Example:
** Item Label: Disp By: ***: Dog 21.00 x Clinacin
   Tabs 150Mg One \& half (1.5) tablets twice a
   day with food ***
** Item Name: Clinacin Tabs 150Mg (100) Clindamycin
** Weight: 30kgs
** Medication Unit Size: 150.0
** Units Dispensed: 21.0
** Dose: 7.5

** Example:
** Item Label: PM:Disp:***: Cat 7.00 x Baytril 50Mg
   Tab Half (1/2) tablet once a day Give until
   finished. ***
** Item Name: Baytril  50Mg Tab (100)  (enrofloxacin)
** Weight: 5kgs
** Medication Unit Size: 50.0
** Units Dispensed: 7.0
** Dose: 5

** Example:
** Item Label: Vet: ***: *** : Dog 10.00 x Veraflox
   Dog 60mg Give ONE (1) tablet ONCE a day Give
   until finished; ***
** Item Name: Veraflox Dog 60mg (70) (pradofloxacin)
** Weight: 30kgs
** Medication Unit Size: 60.0
** Units Dispensed: 10.0
** Dose: 2

** Instance to Label:
** Item Label: Vet: ***: Dog 10.00 x Clavulox Tabs
   500Mg One (1) tablet twice a day Give until
   finished. with food ***
** Item Name: Clavulox (Clavulanic Acid) Tabs 500Mg
   (100) **\\ Weight: 29kgs
** Medication Unit Size: 500.0
** Units Dispensed: 10.0
** Dose:
```

A.2 Prompts Used

This section details the example prompts used for each label in the task.

- **Active Ingredient:** "Referencing the trade name, choose the active ingredient from the Ingredients List that forms the medication. For combination drugs, ensure to select the ingredient with all components."
- **Clinical Indication:** "Using the provided list of possible indications, give the most likely indication for the antimicrobial administration. If unclear from the text, label as unknown."
- **Frequency:** "How many times per day is the medication given?"
- **Medication Size:** "What is the medication unit size in mg?"
- **Dosage Unit Size:** "How many units of the medication are given per dose?"
- **Overall Dose:** "Output the dosage in mg/kg. Dose is determined by multiplying the total dose units given per administration multiplied by the size of the tablet in mg, dividing by the

	Ingredient	*Indication*	*Dose*	*Duration*	*Frequency*	*Dose Unit Size*	*Weight*	*Total Units*	*Medication Size*
RxVetBERT	100	80.0	89.1	88.0	97.0	89.0	-	-	-
Few-Shot Similar Examples									
GPT-3.5	97.5	56.5	29.5	70.0	98.0	98.0	90.5	100	100
GPT-4	99.5	75.0	85.0	91.0	98.5	99.5	100	99.5	100
LLAMA2-70B	94.0	9.0	12.5	58.0	97.5	95.5	100	100	100
Few-Shot Random Examples									
GPT-3.5	67.0	73.5	26.0	61.0	98.5	97.0	99.5	100	100
GPT-4	100	73.5	88.5	84.5	98.5	100	100	99.5	100
LLAMA2-70B	42.0	27.5	9.5	61.0	97.5	92.5	100	100	100
Zero-Shot									
GPT-3.5	80.5	35.0	3.5	52.5	12.0	21.0	69.5	94	100
GPT-4	97.5	69.5	24.0	75.5	97.5	55.0	98.5	100	100
LLAMA2-70B	21.0	0.0	5.0	57.5	98.0	59.5	92.0	99.5	100

Table 3: Accuracy (%) of Large Language Models (LLMs) across multiple settings for all prescription elements, benchmarked against the RxVetBERT baseline ensemble methods.

weight of the patient in kg to determine the mg per kg."

- **Treatment Duration:** "Calculate the length of administration (in days) for the given prescription. To determine the length of administration, find the total number of tablets or doses dispensed and divide by the number of doses given per day."

A.3 Additional Evaluations

While evaluations were performed on all aspects of the prescription text, we omitted the performance of the elements which could be extracted directly out of the text, which were required for the dose calculations but did not require any conversion into numerical values, this included the Weight, Total Units, or Medication Size. We have included this in Table 3.