Automated Evaluation of Antibiotic Prescribing Guideline Concordance in Pediatric Sinusitis Clinical Notes

Davy Weissenbacher[†], PhD

Department of Computational Biomedicine, Cedars-Sinai Medical Center,
Los Angeles, CA, USA
E-mail: davy.weissenbacher@cshs.org

Lauren Dutcher[†], MD, MSCE

Division of Infectious Diseases, Department of Medicine, University of Pennsylvania Perelman School of Medicine,
Philadelphia, PA, USA
E-mail: LDutcher@pennmedicine.upenn.edu

Mickael Boustany, MD

Division of Infectious Diseases, Children's Hospital of Philadelphia,
Philadelphia, PA, USA

Leigh Cressman, MA

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine,
Philadelphia, PA, USA
E-mail: crel@pennmedicine.upenn.edu

Karen O'Connor, MS

Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine,
Philadelphia, PA, USA
E-mail: karoc@pennmedicine.upenn.edu

Keith W. Hamilton, MD

Division of Infectious Diseases, Department of Medicine, University of Pennsylvania Perelman School of Medicine,
Philadelphia, PA, USA
E-mail: Keith.Hamilton@pennmedicine.upenn.edu

[†] Both authors contributed equally to this work

Jeffrey Gerber, MD, PhD
Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine,
Philadelphia, PA, USA
E-mail: GERBERJ@chop.edu

Robert Grundmeier, MD
Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia,
Philadelphia, PA, USA
E-mail: GRUNDMEIER@chop.edu

Graciela Gonzalez-Hernandez, PhD
Department of Computational Biomedicine, Cedars-Sinai Medical Center,
Los Angeles, CA, USA
E-mail: Graciela.GonzalezHernandez@csmc.edu

Background: Ensuring antibiotics are prescribed only when necessary is crucial for maintaining their effectiveness and is a key focus of public health initiatives worldwide. In cases of sinusitis, among the most common reasons for antibiotic prescriptions in children, healthcare providers must distinguish between bacterial and viral causes based on clinical signs and symptoms. However, due to the overlap between symptoms of acute sinusitis and viral upper respiratory infections, antibiotics are often over-prescribed.
Objectives: Currently, there are no electronic health record (EHR)-based methods, such as lab tests or ICD-10 codes, to retroactively assess the appropriateness of prescriptions for sinusitis, making manual chart reviews the only available method for evaluation, which is time-intensive and not feasible at a large scale. In this study, we propose using natural language processing to automate this assessment.
Methods: We developed, trained, and evaluated generative models to classify the appropriateness of antibiotic prescriptions in 300 clinical notes from pediatric patients with sinusitis seen at a primary care practice in the Children's Hospital of Philadelphia network. We utilized standard prompt engineering techniques, including few-shot learning and chain-of-thought prompting, to refine an initial prompt. Additionally, we employed Parameter-Efficient Fine-Tuning to train a medium-sized generative model Llama 3 70B-instruct.
Results: While parameter-efficient fine-tuning did not enhance performance, the combination of few-shot learning and chain-of-thought prompting proved beneficial. Our best results were achieved using the largest generative model publicly available to date, the Llama 3.1 405B-instruct. On our evaluation set, the model correctly identified 94.7% of the 152 notes where antibiotic prescription was appropriate and 66.2% of the 83 notes where it was not appropriate. However, 15 notes that were insufficiently, vaguely, or ambiguously documented by physicians posed a challenge to our model, as none were accurately classified.
Conclusion: Our generative model demonstrated good performance in the challenging task of chart review. This level of performance may be sufficient for deploying the model within the EHR, where it can assist physicians in real-time to prescribe antibiotics in concordance with the guidelines, or for monitoring antibiotic stewardship on a large scale.

Keywords: Antibiotic Stewardship, Classification, Large Language Models, Generative Systems

1. Introduction

Antibiotic stewardship programs (ASPs) aim to optimize the use of antibiotics for specific conditions and to combat the growing threat of antimicrobial resistance.[1] Inappropriate prescribing of antibiotics not only contributes to a global health crisis but also exposes patients, particularly pediatric patients, to unnecessary side effects and disrupts their healthy microbiota.[2] Ensuring that antibiotics are prescribed adequately —only when necessary and with the correct dosage and duration— is essential for maintaining their efficacy and is a key focus in public health and research efforts at national and international levels.

Most antibiotic prescribing takes place in the ambulatory setting, and approximately 30% of all outpatient antibiotic prescriptions are unnecessary; a majority of unnecessary outpatient prescribing is for acute upper respiratory tract infections.[3,4] In particular, sinusitis which is among the most common reasons for ambulatory antibiotic prescribing in children.[3] The symptoms of acute sinusitis often overlap significantly with those of uncomplicated viral upper respiratory tract infections. As a result, antibiotics are often over-prescribed for sinusitis, despite guidelines recommending more conservative use.[5,6]

The Centers for Disease Control and Prevention (CDC) Core Elements of Outpatient Antibiotic Stewardship recommend tracking and reporting ambulatory antibiotic prescribing.[7] Some metrics using data from the electronic health record (EHR) have been developed in order to measure unnecessary and guideline-discordant prescribing.[8] Several studies have created classification models to assess appropriate antibiotic prescribing by linking patient diagnoses to tier-based rules where the antibiotic prescription is always, sometimes, or never appropriate depending on the diagnosis.[3,9,10] Others have focused on metrics for specific conditions, such as acute bronchitis, or have addressed antibiotic selection or duration of therapy.[11–13] These metrics have successfully been used in feedback for clinicians and practices and in assessing the impact of stewardship programs on prescribing.

However, while these metrics and classification schemes perform reasonably well, they have primarily only used structured data from the EHR, and have not been able to use information from unstructured text present in clinical notes. This creates a significant gap for conditions in which the assessment of appropriateness using an electronically-based metric from structured data is not feasible. For example, in acute sinusitis, healthcare providers must distinguish bacterial from viral sinusitis based on clinical signs and symptoms alone, and antibiotic prescribing is only considered guideline-concordant for bacterial sinusitis. As such, there are no lab tests or ICD-10 codes (structured data) that can be used to retroactively measure prescribing appropriateness in the absence of time-intensive manual chart review of clinical notes. While audits of patient charts have elicited important findings for the field of antibiotic stewardship, there are limitations to manual review.[9,14,15] Retrospective manual review of charts is labor intensive and time consuming, therefore only small samples of charts can be reviewed, limiting the potential applications in large scale antibiotic stewardship interventions.

This paper explores the significance of antibiotic stewardship for pediatric sinusitis and presents a generative system, utilizing a Large Language Model (LLM) approach, to automate the analysis of unstructured notes from pediatric primary care practices to determine justified vs unjustified prescription of antibiotics given a case presentation, seeking to enable a large-scale study that aims to improve prescribing practices.

2. Materials and Methods

We represented the task of evaluating the guideline concordance of antibiotic prescribing in clinical notes as a decision task. That is, given a note in which a patient was diagnosed with sinusitis and prescribed antibiotics, our system should predict whether the prescription was 1) appropriate, 2) not appropriate, or 3) insufficient or ambiguous, in cases where the note does not contain enough information to assess the appropriateness of the prescription.

2.1. *Data collection*

We identified all pediatric (younger than 18) clinical encounter notes by ICD-10 code from outpatient billed encounters at one of 32 primary care practices in the Children's Hospital of Philadelphia (CHOP) network from July 1, 2017 through June 30, 2021 using the following criteria: 1) visits with either a J01 (acute sinusitis) or J32 (chronic sinusitis) code and 2) a prescription of an oral antibiotic (excluding antibiotics that would never be prescribed for sinusitis). The following patients were excluded: 1) patients with a confounding chronic medical condition identified by an ICD-10 code;[16] 2) patients with an ICD-10 code for another infection that would warrant an antibiotic prescription at the same visit. Only primary care visits were included; emergency department and urgent care visits were excluded. Only office visit notes from healthcare providers were included.

A total of 10,311 patients met the inclusion criteria 6,377 (61.9%) for acute sinusitis, and 3,934 (38.2%) for chronic sinusitis, seen by 310 providers. The median number of encounters per provider was 12 (3 – 48). To develop, train, and evaluate our classifier, we selected 300 encounter notes at random. Our intent was to reflect the natural distribution of the notes where the system will be deployed, so we did not oversample or undersample any specific group or provider. This resulted in 190 (63.3 %) encounter notes for acute sinusitis and 110 (36.7%) for chronic sinusitis, seen by 132 providers. The median number of encounters per provider was 50 (21.5 – 92).

We split our annotated dataset into three sets, the first two of which were selected from 80 percent of the providers: a training set with 200 notes (117 notes with appropriate prescriptions, 69 not appropriate, and 14 with insufficient or ambiguous documentation), a development set with 50 notes (32 appropriate, 16 not appropriate and 2 insufficient). For the third set (the test set), we selected 50 notes from the remaining 20% of the providers (35 appropriate, 14 not appropriate, 1 insufficient), in order to be able to test the system on how it adapts to notes from new (unseen) providers.

2.2. *Annotation*

We derived a set of criteria by adapting the recommendations of two clinical practice guidelines[17,18] to define the appropriateness of antibiotic prescribing to the patients we selected. Table 1 summarizes our criteria. If a patient met at least one criterion, our annotators labeled the note as appropriate. If there was clear evidence in the note that none of the criteria were met, the annotators labeled the note inappropriate; otherwise, if it was not possible for the annotator to decide if the criteria were met or not in a note, the note was labeled insufficient. Such cases usually include incomplete, ambiguous or vague documentation. The phrase ”*patient had congestion for over a week*” is an example of an ambiguous documentation. If the congestion lasted for 8 or 9 days, criterion 1 in Table 1 would not be met and this would be labeled ’not justified’. However, if the symptom lasted 10 days or longer, then criterion 1 would be satisfied and this would be labeled ’justified’. The phrase ”*Fever x 3 days*” is an example of incomplete documentation because it does not specify the exact temperature. Note that our definition focuses solely on the act of prescribing antibiotics and excludes considerations related to the appropriateness of the specific antibiotic prescribed, as well as its dosage and duration.

Table 1: Clinical guidelines used to assess the appropriateness of an antibiotic prescription for patients diagnosed with sinusitis. If the clinical note provided sufficient evidence to meet at least one of the three established criteria, the prescription was annotated as appropriate.

Antibiotics appropriateness
1. Persistent illness: nasal discharge (of any quality), daytime cough, or sinus pain/pressure lasting for $\geq$ 10 days without improvement
2. Severe onset, i.e., concurrent fever (temperature $\geq$ 39°C/102.2°F) and purulent nasal discharge or sinus pain/pressure for at least 3 consecutive days
3. Worsening course, i.e., worsening or new onset of nasal discharge, daytime cough, sinus pain/pressure, or fever after initial improvement

One pediatric physician annotated the 300 notes of our corpus as *appropriate*, *inappropriate*, or *insufficient*. A second pediatric physician is currently annotating 50 notes of our corpus to compute the inter-annotator agreement. To guide this assessment, an annotation guide was developed, using input from a primary care pediatrician, two infectious diseases specialists, and one pediatric infectious diseases specialist. This annotation guide was developed iteratively using practice notes from the same practices with the goal of improving reproducibility as much as possible.

2.3. *Generative models*

Our task presents a significant challenge for conventional natural language processing (NLP) systems, which typically rely on a pipeline approach.[19–21] In such systems, a task is divided into several ’simpler’ subtasks, each performed sequentially by independent modules. To complete our task, an NLP pipeline would first require an information extraction module to identify key symptoms in the clinical notes —congestion, cough, sinus pain/discomfort, and fever— as reported by the patient during the encounter. Next, a classification module would detect mentions of symptom severity and assign appropriate labels to each symptom. A third module would normalize the extracted information by identifying and representing the progression of symptoms. Finally, a logical validation module would verify whether the extracted information aligns with the criteria outlined in Table 1, ultimately generating the final decision.

This pipeline approach has several limitations that often result in reduced performance[22] and limited adoption in the medical field. Each module operates based on a set of rules, which can either be manually crafted or automatically learned from training data. Both approaches demand significant human effort. In the medical domain, writing rules requires expertise in both computer science and medicine, and these rules are often difficult to maintain over time.[23] An alternative is to learn the rules directly from annotated examples,[24] but standard machine learning algorithms typically need thousands of examples to achieve acceptable performance, a resource-intensive and costly process. This often leaves modules only partially trained, leading to suboptimal results.[25] Even when the rules are well defined, they rarely account for all possible cases, and module performance is almost never flawless.[24] Since a pipeline approach processes data sequentially through imperfect modules, errors from earlier stages propagate through the system, compounding in later stages and significantly limiting overall performance.[19,26] Moreover, conventional NLP modules —such as classifiers, sequence labelers, or normalizers— are typically designed to output only their labels and confidence scores, without providing explanations for their decisions. This lack of interpretability forces experts to rely on ad-hoc algorithms producing only partial and incomplete explanation of the module behavior.[27] This issue was particularly pronounced with transformer-based encoders like BERT,[28,29] the standard NLP architecture before the recent advancements with large language models-based generative models, which was often qualified as a black box system and not well adopted by medical professionals who doubted their decision.

As an alternative to conventional NLP systems, we propose using state-of-the-art generative systems powered by large language models (LLMs). In recent years, generative systems have become the leading approach in NLP as evidenced by the widespread success of chatGPT.[30] Generative systems feature interfaces that allow users to submit prompts in natural language, an intuitive interface to perform a task.[31] These prompts typically include an instruction specifying the desired action, along with optional data needed to perform the task. Generative systems leverage semi-supervised training to transfer general knowledge acquired from extensive text corpora, enabling them to generate appropriate responses and execute instructions for tasks they were not explicitly trained on. This eliminates the need to retrain the system for each specific task, a requirement often necessary in conventional NLP systems.

In this study, we applied a generative system to address the specific challenge of antibiotic stewardship, a task for which no established benchmarks exist. In accordance with common practices for deploying generative systems in clinical settings,[32] we utilized prompt engineering with few-shot learning and chain-of-thought reasoning. Instead of adopting more advanced and resource-intensive techniques —such as full fine-tuning on large clinical datasets,[33] knowledge injection via retrieval-augmented generation,[34] or self-correction through multi-agent interactions— we chose to evaluate the system's inherent capabilities,[35,36] reserving these enhancements for future research.

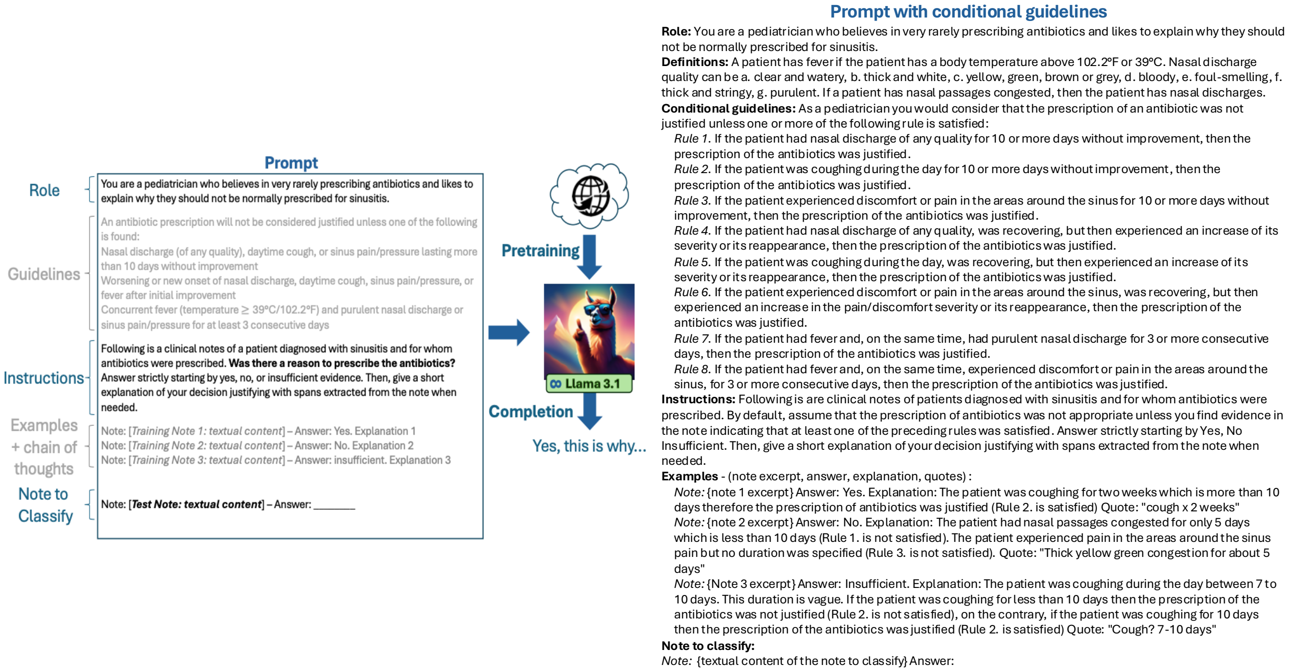
2.4. *Classification with Generative systems*

We performed our classification using generative systems from the Llama 3 family,[37] which is one of the largest freely available sets of models offering competitive performance compared to proprietary alternatives. We progressively refined an initial simple prompt by following few-shot learning and chain-of-thought techniques to enhance the models' performance on our task. Additionally, we fine-tuned a Llama-3-70B-Instruct model using a parameter-efficient fine-tuning (PEFT) approach, namely LoRA,[38] to specialize the model for our specific task.

Initial prompt. Figure 1 outlines the various components of the prompt we designed to instruct our model on how to classify the appropriateness of antibiotic prescription in clinical notes. We began our experiments with an initial straightforward prompt that defined the role the generative model should assume, followed by a brief paragraph specifying the instructions for the task. This paragraph included the following key components:

(1) The role specifying the function the model should adopt when generating response, in our case a pediatrician.
(2) The context which describes the notes; specifically, the input note is a clinical note of a patient diagnosed with sinusitis who received antibiotics.
(3) The question the model should answer.
(4) The format in which we wanted the model to present its response.
(5) The text of the note to be classified
(6) The keyword *Answer:* to initiate the model's completion according to our specified format.

The authors, during an interactive session, tried multiple initial prompts and evaluated the Llama 3 70B-instruct model's results on the development set. At the end of the interactive session, we selected the initial prompt illustrated in Figure 1 Left.



You are a pediatrician who believes in very rarely prescribing antibiotics and likes to explain why they should not be normally prescribed for sinusitis.

An antibiotic prescription will not be considered justified unless one of the following is found:
Nasal discharge (of any quality), daytime cough, or sinus pain/pressure lasting more than 10 days without improvement
Worsening or new onset of nasal discharge, daytime cough, sinus pain/pressure, or fever after initial improvement
Concurrent fever (temperature ≥ 39°C/102.2°F) and purulent nasal discharge or sinus pain/pressure for at least 3 consecutive days

Following is a clinical notes of a patient diagnosed with sinusitis and for whom antibiotics were prescribed. **Was there a reason to prescribe the antibiotics?** Answer strictly starting by yes, no, or insufficient evidence. Then, give a short explanation of your decision justifying with spans extracted from the note when needed.

Note: [*Training Note 1: textual content*] – Answer: Yes. Explanation 1
Note: [*Training Note 2: textual content*] – Answer: No. Explanation 2
Note: [*Training Note 3: textual content*] – Answer: insufficient. Explanation 3

Note: [***Test Note: textual content***] – Answer: ________

Prompt with conditional guidelines

Role: You are a pediatrician who believes in very rarely prescribing antibiotics and likes to explain why they should not be normally prescribed for sinusitis.
Definitions: A patient has fever if the patient has a body temperature above 102.2°F or 39°C. Nasal discharge quality can be a. clear and watery, b. thick and white, c. yellow, green, brown or grey, d. bloody, e. foul-smelling, f. thick and stringy, g. purulent. If a patient has nasal passages congested, then the patient has nasal discharges.
Conditional guidelines: As a pediatrician you would consider that the prescription of an antibiotic was not justified unless one or more of the following rule is satisfied:
Rule 1. If the patient had nasal discharge of any quality for 10 or more days without improvement, then the prescription of the antibiotics was justified.
Rule 2. If the patient was coughing during the day for 10 or more days without improvement, then the prescription of the antibiotics was justified.
Rule 3. If the patient experienced discomfort or pain in the areas around the sinus for 10 or more days without improvement, then the prescription of the antibiotics was justified.
Rule 4. If the patient had nasal discharge of any quality, was recovering, but then experienced an increase of its severity or its reappearance, then the prescription of the antibiotics was justified.
Rule 5. If the patient was coughing during the day, was recovering, but then experienced an increase of its severity or its reappearance, then the prescription of the antibiotics was justified.
Rule 6. If the patient experienced discomfort or pain in the areas around the sinus, was recovering, but then experienced an increase in the pain/discomfort severity or its reappearance, then the prescription of the antibiotics was justified.
Rule 7. If the patient had fever and, on the same time, had purulent nasal discharge for 3 or more consecutive days, then the prescription of the antibiotics was justified.
Rule 8. If the patient had fever and, on the same time, experienced discomfort or pain in the areas around the sinus, for 3 or more consecutive days, then the prescription of the antibiotics was justified.
Instructions: Following is are clinical notes of patients diagnosed with sinusitis and for whom antibiotics were prescribed. By default, assume that the prescription of antibiotics was not appropriate unless you find evidence in the note indicating that at least one of the preceding rules was satisfied. Answer strictly starting by Yes, No Insufficient. Then, give a short explanation of your decision justifying with spans extracted from the note when needed.
Examples - (note excerpt, answer, explanation, quotes) :
Note: {note 1 excerpt} Answer: Yes. Explanation: The patient was coughing for two weeks which is more than 10 days therefore the prescription of antibiotics was justified (Rule 2. is satisfied) Quote: "cough x 2 weeks"
Note: {note 2 excerpt} Answer: No. Explanation: The patient had nasal passages congested for only 5 days which is less than 10 days (Rule 1. is not satisfied). The patient experienced pain in the areas around the sinus pain but no duration was specified (Rule 3. is not satisfied). Quote: "Thick yellow green congestion for about 5 days"
Note: {Note 3 excerpt} Answer: Insufficient. Explanation: The patient was coughing during the day between 7 to 10 days. This duration is vague. If the patient was coughing for less than 10 days then the prescription of the antibiotics was not justified (Rule 2. is not satisfied), on the contrary, if the patient was coughing for 10 days then the prescription of the antibiotics was justified (Rule 2. is satisfied) Quote: "Cough? 7-10 days"
Note to classify:
Note: {textual content of the note to classify} Answer:

Fig. 1: Left: Iterative construction of a prompt to classify antibiotic prescription appropriateness using a Llama 3 generative model. Right: Our prompt with conditional guidelines.

Guidelines. We first extended this initial prompt by inserting the clinical guidelines that our annotators followed when labeling the notes in our corpus. The generative models that are publicly available were pretrained on large corpora from the internet, which contain few, if any, professional medical documents[39,40] and may not have encountered or memorized our specific guidelines. By including these guidelines directly in the prompt, we ensured that the model had direct access to the criteria defining the task. The guidelines in Table 1 were written for medical professionals. To make them more accessible and easier for our generative model to interpret, *LD*, Assistant Professor in Medicine, simplified the language used in the guidelines.

Few-shot learning. Much like humans, generative models can benefit from seeing a few examples before attempting a task, a concept known as few-shot learning. We implemented this approach by including the text of three notes from our training set in the prompt, each accompanied by their appropriateness labels. Although complex conditions could be used to select these examples -such as choosing notes with close semantic similarity to the one being classified or those that annotators found challenging[41]- we opted to select the examples randomly. We chose this approach for simplicity and left the exploration of more sophisticated selection strategies for future work.

Chain-of-thought prompting. Together with few-shot learning, we also employed chain-of-thought prompting.[42] After each label of our training examples, we included a brief explanation of the label, along with the relevant quotes that demonstrated the extracted span from the example note supporting the explanation. *LD* provided these explanations, highlighting which criteria from Table 1 were met, missing, or challenging to verify based solely on the note's text. Requesting explanations along with quotes forces the model to ground its responses within the text of the notes, thereby reducing hallucinations. Despite the large context window of 8,192 tokens, the Llama-3-70B-Instruct model still has a limited prompt capacity, which restricted us to including no more than three example notes. In a supplementary experiment, to include additional examples, we did not input the entire text of the training notes. Instead, we truncated the notes, only incorporating the sentences containing the relevant quoted phrases.

We hypothesized that chain-of-thought reasoning is an important component for improving the performance of a generative model, and conducted additional experiments by reformulating our initial prompt and its components. While the description of the model's role remained unchanged, we revised the context and question to predispose the model to answer 'not appropriate' by default unless it identified evidence in the notes that satisfied a criterion from our guidelines. We also introduced simple definitions for *'fever,' 'nasal discharge,' and 'nasal congestion'* before presenting the guidelines, and we rephrased the guidelines as a set of eight conditional rules. Additionally, we revised the explanations for all ten examples in the prompt to explicitly indicate which rules were met or unmet (Line 9 in Table 2). We provide the exact prompt used in these experiments in Figure 1) Right.

Parameter Efficient Fine-tuning with LoRA. Although generative models achieve state-of-the-art performance on general NLP tasks, they may benefit from being fine-tuned to perform more specific and challenging tasks. A standard method for training generative models is full fine-tuning, a supervised training process. In this process, the model is presented with instructions and corresponding data required to perform a task. It generates a response, which is then automatically compared to the expected gold-standard answer from the training examples. If the model generates the expected answer, no adjustments are made. However, if it deviates, all weights of its underlying neural network are updated to increase the likelihood of generating the correct response. While fine-tuning can enhance the model's performance on specific tasks, updating all weights of a very large neural network is computationally intensive and requires a significant number of expensive GPUs, which were not available for our experiments. LoRA is a heuristic proposed by Yu et al.[38] to update only a small portion of the weights in the neural network. We trained the Llama 3 70B-instruct model using the implementation of LoRA from the litGPT 0.40 library.[43] We employed the default learning parameters provided in litgpt, which included the cross-entropy loss function and the AdamW optimizer instantiated with a a learning rate equal to 1e-3. The model was trained using bfloat16 precision, with the low-rank adaptation (LoRA) matrix rank set to 32. Training was conducted across 4 A100 GPUs for 20 epochs, with a batch size of 4. We retained the model checkpoint that achieved the highest performance on our development set as the final trained model.

Larger language model. It has been demonstrated that increasing the size of generative models not only improve their performance on known tasks but also unlocks new capabilities exclusive to the largest models.[40] For instance chain of thoughts, sufficiently large models can mimic the logical steps humans follow when solving problems, and by learning to explain their reasoning, they improve their performance. Considering the potential benefits of larger foundational models, we also evaluated the Llama 3.1 405B-instruct quantized (int4) model, which was released shortly before our submission deadline.

Evaluation. We evaluated the performance of the Llama 3 70B-instruct model using our initial prompt on the development set, then assessed its performance as we sequentially added each component designed to enhance the prompt —namely, guidelines, a few examples, explanations for the labels, and finally, fine-tuning on our training corpus. We conducted all experiments with a temperature of 0.001, top-p of 0.01, and top-k of 1, to ensure deterministic responses by consistently selecting the most likely token when generating its answers. Due to time constraints and the slow processing speed of the Llama 3.1 405B-instruct model, approximately 30 minutes to classify a single note, we were unable to rerun all experiments to find the best prompt settings for this model. Instead, we evaluated this model with the best-performing settings from the Llama 3 70B-instruct model on the development set. We conducted all experiments with the default temperature of the Llama 3.1 405B-instruct model set to 0.6, top-p to 0.9, and top-k to 50, allowing for more variety in its responses. Because of time limitations, we did not run additional experiments with the temperature settings adjusted to ensure deterministic responses. After identifying the best model and settings, we performed a final evaluation on the test set. Our test set consists of only 50 notes, making it relatively small. Since our results indicate that the best-performing model used few-shot prompting and was not fine-tuned on our training examples, those examples remained unused. To assess how well our system scales, we reassigned the training examples and evaluated the model on the training set. We define the evaluation set as the combined set of all examples from both the training and test sets. Since there were very few notes labeled as insufficient in our gold standard, most errors involved the model confusing notes with appropriate (guideline-concordant) prescriptions with those that were not appropriate (not guideline-concordant) and vice versa. Therefore, we chose to report all results by only providing the percentage of notes in each class that were correctly labeled by the generative model and did not report the more standard F1-scores.

3. Results and Discussion

We present our results in Table 2, highlighting best performance on the test set. The system correctly identified 10 out of 14 (71.4%) notes labeled as not appropriate and 32 out of 35 (91.4%) notes labeled as appropriate (line 9). This performance was achieved by providing the model with instructions and logical guidelines to perform the task, without training on the training set. The classifier demonstrated good correctness on a complex task typically performed by trained physicians when only given a few examples and clear explanations indicating whether the rules of the guidelines were satisfied or not. We also evaluated the model's performance on the entire training set to provide a more comprehensive assessment (line 9). On this larger dataset, the system maintained comparable performance, with improved detection of notes labeled as appropriate, correctly identifying 112 out of 117 (95.7%), while its detection of unlabeled notes was slightly lower, identifying 45 out of 69 (65.2%).

The table shows that all modifications made to the initial prompt (line 1.) led to incremental improvements in the model's classification performance. The table offers several interesting insights. Firstly, it is surprising that truncating the text of the example notes did not lead to a performance drop (line 4. *vs.* line 5.). This suggests that most of the text in a note is not utilized by the model for understanding the examples and can be omitted without losing essential information. Secondly, it is worth noting that LoRA, the parameter-efficient technique we employed using our training set, did not enhance the model's performance (line 6. vs. line 7.).This unexpected result requires additional experiments for further explanation. Thirdly, our findings align with recent trends in the NLP community, which indicate that generative models based on larger language models perform better than their smaller counterparts. This is evident in Table 2, where the Llama 3.1 405B-instruct model, at the time of writing, the largest model freely available to the community, outperformed the Llama 3 70B-instruct model. Lastly, since the model was not trained on our training set, it was not biased toward recognizing the style of certain providers over others. It demonstrated robustness to variations in providers' styles and achieved comparable performance on the test set as it did on the development set.

Several prior studies have used NLP and/or LLMs in infectious diseases to aid in the diagnosis and treatment of infections, such as through the review of radiology reports or in infection surveillance.[44,45] To our knowledge, however, we present the first use of LLMs in the assessment of antibiotic prescribing appropriateness using clinician notes. While these methods require further refinement and validation in larger cohorts, use of LLMs can complement previously-developed EHR-based stewardship metrics that use structured data elements, and thus improve the ability to assess prescribing practices.[3]

The methods presented here have the potential for broad application. Sinusitis is one of the most common infectious diagnoses in the outpatient setting for both adults and children, with

Table 2: Classification results on the evaluation sets for various prompts and large language models. For each line, the top percentage represents the proportion of notes labeled as appropriate correctly retrieved by the model, while the bottom percentage indicates the proportion of notes labeled as not appropriate retrieved. We omitted the percentages of notes labeled as insufficient because none of the models were able to retrieve any in this category.

	Development	**Test**	**Train**
Llama 3 70B-instruct model			
1. Role + Instructions	0.0 (0/32) 1.0 (16/16)	—	—
2. Role + Instructions + Guidelines	0.0 (0/32) 1.0 (16/16)	—	—
3. Role + Instructions + Guidelines + 3 Examples (Full text) w/o explanations	9.4 (3/32) 87.5 (14/16)	—	—
4. Role + Instructions + Guidelines + 3 Examples (Full text) & explanations	87.5 (28/32) 18.8 (3/16)	—	—
5. Role + Instructions + Guidelines + 3 Examples (Excerpt text) & explanations	53.1 (17/32) 62.5 (10/16)	—	—
6. Role + Instructions + Guidelines + 10 Examples (Excerpt text) & explanations	90.6 (29/32) 31.2 (5/16)	—	—
7. Role + Instructions + Guidelines + 10 Examples (Excerpt text) & explanations + LoRA fine-tuning	90.6 (29/32) 31.2 (5/16)	—	—
Llama 3.1 405B-instruct model			
8. Role + Instructions + Guidelines + 10 Examples (Excerpt text) & explanations	93.8 (30/32) 68.8 (11/16)	91.4 (32/35) 64.3 (9/14)	—
9. Role + Instructions + conditional Guidelines + 10 Examples (Excerpt text) & explanations	**90.6** (29/32) **93.8** (15/16)	**91.4** (32/35) **71.4** (10/14)	95.7 (112/117) 65.2 (45/69)

most encounters occurring in primary care, urgent care, and the emergency department.[3] If deployed across these settings, use of LLMs to assess prescribing appropriateness for sinusitis has the potential to impact the care of millions of people. In practice, we envision that this tool could be used in several ways. First, it could be used in provider-based feedback interventions where prescribers receive feedback on their prescribing appropriateness retrospectively at regular intervals (e.g. monthly), similar to prior work that utilized structured EHR-based metrics.[11,12,46] Additionally, this approach could be an important tool in tracking guideline concordant prescribing over time at clinic or health system level, as recommended by the CDC.[7] Finally, this also has the potential to be deployed to aid in real-time decision support during clinic visits, though modifications may need to be made given that not all notes are completed during the visit.

3.1. *Error analysis*

We analyzed the errors made by the best classifier, the Llama 3.1 405B-instruct model (line 8. in Table 2), on the examples in the test set. The model misclassified a total of 8 notes. The most frequent errors were False Positives (FPs), where the notes were labeled as inappropriate for antibiotic prescription, but the classifier predicted them as appropriate. There were 5 such misclassified notes. Upon re-examining the notes, *LD* reviewed the explanations provided by the model and determined whether they were valid. It was found that 2 FPs occurred in notes that could have been labeled as insufficient due to ambiguous temperature documentation. For the remaining 3 FPs, *LD* confirmed the errors made by the system. One error resulted from the incorrect resolution of a deictic time reference; another from a misinterpretation of the term 'worsening' (in the phrase 'acutely worsening symptoms overnight', where 'worsening' refers to an increase in the severity of symptoms, not the progression pattern where the patient initially feels sick, then slightly better, and then worse); and the final FP was due to the system's hallucination, incorrectly stating that a temperature of 102°F is higher than 102.2°F.

The model had more success classifying the notes in which antibiotics were prescribed appropriately. There were only 3 False Negatives (FNs), as these notes clearly mentioned the onset and duration of symptoms. One FN occurred due to the under-specification of the definition of fever in criterion 3 in Table 1; unlike criterion 2, the exact temperature defining a fever is not specified. As a result, there was a disagreement between the annotator and the system regarding the resolution of this criterion in the note. The last two FNs were made on notes that were ambiguous and could have been labeled insufficient.

Finally, we analyzed the errors made by the best-performing classifier (line 9.) on the 15 notes labeled as *insufficient* in the evaluation set. All misclassifications involved ambiguous symptom duration phrases, such as "*congestion for over a week*", "*cough 7-10 days*", or "*nasal discharge about 1.5 weeks*". In 7 instances, the model correctly identified the temporal expressions that were vague but failed to recognize the ambiguity and inaccurately assign either a shorter or longer duration. In 8 other cases, the model explicitly flagged the expressions as ambiguous but it still opted for an incorrect duration inference. Given that all errors stemmed from ambiguity in symptom duration —often involving similar phrasing— we could incorporate additional examples into the prompt to help the model better recognize those phrases and correctly class insufficient documentation.

3.2. *Limitations and future work*

The largest model, Llama 3.1 405B-instruct, demonstrated good performance on our task. It was able to follow the logic of our guidelines and provide reasonable explanations for its decisions without explicit training. Although the task is challenging, it only requires the system to identify four common symptoms, assess their severity, and understand their duration or progression patterns. As evidenced by our performance with general generative models, the necessary knowledge to perform the task was available in their training data from the internet. However, most clinical NLP tasks will require specialized knowledge available only in clinical notes and ontologies. Researchers will need to continue pretraining or fine-tuning these models to integrate this domain-specific knowledge. As the size of generative models continues to grow, these training tasks become increasingly challenging for standard institutions such as hospitals and universities, which may lack the necessary hardware for the required computations.[47]

Note that our evaluation has several limitations. First, all notes were sampled from a single clinical institution. We are currently annotating 281 notes from primary care visits for adult sinusitis at one of the University of Pennsylvania Health System's practices. To assess the robustness of our system in a different clinical setting, we plan to apply it to these newly annotated notes. Additionally, future evaluations should test the accuracy of our methods in other clinical environments, such as urgent care and emergency departments, and across institutions that use different EHR systems. Second, our cohort was identified using ICD-10 codes, which have suboptimal sensitivity and specificity for infectious diagnoses.[48] Moreover, we only included visits where an antibiotic was prescribed. It is possible that some visits for sinusitis, where an antibiotic was justified but not prescribed (guideline discordant), were missed. However, given that the majority of patients with a sinusitis diagnosis receive antibiotics, this scenario is likely infrequent.[46]

Ambiguous and vague documentation in the notes continues to pose a challenge for our best model, as none of the insufficiently documented notes were correctly classified. With larger language models now supporting input prompts of up to 16,000 tokens, we plan to include more examples of vague and ambiguous notes, along with explanations, to help the model recognize and classify these cases appropriately. Despite forcing the models to justify their decisions and anchor their answers within the input texts, we still found instances of hallucination. Integrating 'debates' among several generative LLM-based models has been proposed as an effective solution to detect and reduce hallucinations.[35,49] Our approach could easily be extended from a single generative model performing classification to a deliberative panel finding consensus for each debated note. We leave the deployment and evaluation of this approach to future work.

4. Conclusion

To address the challenge of over-prescribing antibiotics for sinusitis in children, this study proposes using natural language processing to automate the assessment of prescription appropriateness, overcoming the limitations of time-consuming manual chart reviews. We developed, trained, and evaluated generative models to classify the appropriateness of antibiotic prescriptions in 300 clinical notes from pediatric patients with sinusitis at the Children's Hospital of Philadelphia primary care network. Although Parameter-Efficient Fine-Tuning did not improve performance, the combination of few-shot learning and chain-of-thought prompting proved beneficial. Our best results were achieved using the largest generative model available at the time, the Llama 3.1 405B-instruct. On our evaluation set, the model correctly identified 144 (94.7%) of the 152 notes where the antibiotic prescription was appropriate and 55 (66.2%) of the 83 notes where it was not. Without training, our generative model demonstrated good performance in this complex task, suggesting it could be effectively deployed within the EHR to assist physicians in real-time to prevent over-prescribing as well as in monitoring antibiotic prescribing on a large scale. The clinical notes annotated for this study are Protected Health Information and not publicly available at this point. We have shared the code for access at `https://bitbucket.org/hlpgonzalezlab/naps/`.

Acknowledgements

The research reported in this publication was supported by a Centers for Disease Control and Prevention (CDC) Cooperative Agreement Funding Opportunity Announcement (FOA) U54 CK000610, Epicenters for the Prevention of Healthcare-Associated Infections.

References

1. T. F. Barlam, S. E. Cosgrove, L. M. Abbo *et al.*, Implementing an Antibiotic Stewardship Program: Guidelines by the Infectious Diseases Society of America and the Society for Healthcare Epidemiology of America, *Clinical Infectious Diseases* **62**, e51 (04 2016).
2. L. McDonnell, A. Gilkes, M. Ashworth, V. Rowland, T. H. Harries, D. Armstrong and P. White, Association between antibiotics and gut microbiome dysbiosis in children: systematic review and meta-analysis, *Gut Microbes* **13**, p. 1870402 (2021).
3. K. E. Fleming-Dutra, A. L. Hersh, D. J. Shapiro *et al.*, Prevalence of Inappropriate Antibiotic Prescriptions Among US Ambulatory Care Visits, 2010-2011, *JAMA* **315**, 1864 (05 2016).
4. K. J. Suda, L. A. Hicks, R. M. Roberts, R. J. Hunkler, L. M. Matusiak and G. T. Schumock, Antibiotic Expenditures by Medication, Class, and Healthcare Setting in the United States, 2010–2015, *Clinical Infectious Diseases* **66**, 185 (08 2017).
5. A. W. Chow, M. S. Benninger, I. Brook, J. L. Brozek, E. J. C. Goldstein, L. A. Hicks, G. A. Pankey, M. Seleznick, G. Volturo, E. R. Wald and J. File, Thomas M., IDSA Clinical Practice Guideline for Acute Bacterial Rhinosinusitis in Children and Adults, *Clinical Infectious Diseases* **54**, e72 (04 2012).
6. E. R. Wald, K. E. Applegate, C. Bordley, D. H. Darrow, M. P. Glode, S. M. Marcy, C. E. Nelson, R. M. Rosenfeld, N. Shaikh, M. J. Smith, P. V. Williams and S. T. Weinberg, Clinical Practice Guideline for the Diagnosis and Management of Acute Bacterial Sinusitis in Children Aged 1 to 18 Years, *Pediatrics* **132**, e262 (07 2013).
7. Core elements of outpatient antibiotic stewardship `https://www.cdc.gov/antibiotic-use/hcp/core-elements/outpatient-antibiotic-stewardship.html`, Accessed July 30, 2024.
8. Measurement and evaluation approaches to improve outpatient antibiotic prescribing in health systems `https://www.cdc.gov/antibiotic-use/pdfs/Measurement-Evaluation-Improve-Outpatient-508.pdf`, Accessed July 30, 2024.
9. K. O. Degnan, V. Cluzet, M. Z. David, L. Dutcher, L. Cressman, E. Lautenbach and K. W. Hamilton, Development and validation of antibiotic stewardship metrics for outpatient respiratory tract diagnoses and association of provider characteristics with inappropriate prescribing, *Infection Control 38; Hospital Epidemiology* **43**, p. 56–63 (2022).
10. K.-P. Chua, M. A. Fischer and J. A. Linder, Appropriateness of outpatient antibiotic prescribing among privately insured US patients: ICD-10-CM based cross sectional study, **364**, p. k5092.
11. D. Meeker, J. A. Linder, C. R. Fox, M. W. Friedberg, S. D. Persell, N. J. Goldstein, T. K. Knight, J. W. Hay and J. N. Doctor, Effect of Behavioral Interventions on Inappropriate Antibiotic Prescribing Among Primary Care Practices: A Randomized Clinical Trial, *JAMA* **315**, 562 (02 2016).
12. J. S. Gerber, P. A. Prasad, A. G. Fiks, A. R. Localio, R. W. Grundmeier, L. M. Bell, R. C. Wasserman, R. Keren and T. E. Zaoutis, Effect of an Outpatient Antimicrobial Stewardship Intervention on Broad-Spectrum Antibiotic Prescribing by Primary Care Pediatricians: A Randomized Trial, *JAMA* **309**, 2345 (2013).
13. L. Dutcher, K. Degnan, A. B. Adu-Gyamfi, E. Lautenbach, L. Cressman, M. Z. David, V. Cluzet, J. E. Szymczak, D. A. Pegues, W. Bilker, P. Tolomeo, f. t. C. f. D. C. Hamilton, Keith W and P. C. P. E. Program, Improving Outpatient Antibiotic Prescribing for Respiratory Tract Infections in

Primary Care: A Stepped-Wedge Cluster Randomized Trial, *Clinical Infectious Diseases* **74**, 947 (07 2021).

14. K. N. Truitt, T. Brown, J. Y. Lee and J. A. Linder, Appropriateness of antibiotic prescribing for acute sinusitis in primary care: A cross-sectional study, **72**, 311.
15. D. J. Livorsi, C. M. Linn, B. Alexander, B. H. Heintz, T. A. Tubbs and E. N. Perencevich, The value of electronically extracted data for auditing outpatient antimicrobial prescribing, **39**, 64.
16. C. Feudtner, J. A. Feinstein, W. Zhong, M. Hall and D. Dai, Pediatric complex chronic conditions classification system version 2: updated for icd-10 and complex medical technology dependence and transplantation, *BMC Pediatrics* **14**, p. 199 (2014).
17. A. Chow, M. Benninger, I. Brook *et al.*, Idsa clinical practice guideline for acute bacterial rhinosinusitis in children and adults, *Clin Infect Dis* **54**, e72 (2012).
18. E. R. Wald, K. E. Applegate, C. Bordley *et al.*, Clinical Practice Guideline for the Diagnosis and Management of Acute Bacterial Sinusitis in Children Aged 1 to 18 Years, *Pediatrics* **132**, e262 (07 2013).
19. H. Cunningham, V. Tablan, A. Roberts and K. Bontcheva, Getting more out of biomedical documents with gate's full lifecycle open source text analytics, *PLOS Computational Biology* **9**, 1 (02 2013).
20. C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky, The Stanford CoreNLP natural language processing toolkit, in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014.
21. D. Ferrucci, A. Lally, K. Verspoor and E. Nyberg, Unstructured information management architecture (UIMA) version 1.0 OASIS Standard (mar, 2009).
22. D. Weissenbacher, Influence des annotations imparfaites sur les systèmes de traitement automatique des langues, un cadre applicatif: la résolution de l'anaphore pronominale, PhD thesis, Université Paris-Nord - Paris XIII2008.
23. T. Poibeau, *Extraction Automatique D'information* (Hermes, 2003).
24. A. Magge, D. Weissenbacher, A. Sarker, M. Scotch and G. Gonzalez-Hernandez, Deep neural networks and distant supervision for geographic location mention extraction, *Bioinformatics* **34**, i565 (2018).
25. W. H. Clark and A. J. Michaels, Training from zero: Forecasting of radio frequency machine learning data quantity, *Telecom* **5**, 632 (2024).
26. D. Weissenbacher and Y. Sasaki, Which factors contributes to resolving coreference chains with bayesian networks?, in *14th International Conference on Intelligent Text Processing and Computational Linguistics*, 2013.
27. A. Thampi, *Interpretable AI: : Building explainable machine learning systems* (Manning, 2022).
28. J. Vig, A multiscale visualization of attention in the transformer model, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (Association for Computational Linguistics, July 2019).
29. A. Rogers, O. Kovaleva and A. Rumshisky, A Primer in BERTology: What We Know About How BERT Works, *Transactions of the Association for Computational Linguistics* **8**, 842 (2021).
30. L. Ouyang, J. Wu, X. Jiang *et al.*, Training language models to follow instructions with human feedback, in *Advances in Neural Information Processing Systems*, eds. S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (Curran Associates, Inc., 2022).
31. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.* **21** (jan 2020).
32. R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting and N. Liu, Large language models in health care: Development, applications, and challenges, *Health Care Science* **2**, 255 (2023).

33. X. Yang, A. Chen, N. PourNejatian *et al.*, A large language model for electronic health records, *npj Digit. Med* **5** (2022).
34. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel and D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20 (Curran Associates Inc., Red Hook, NY, USA, 2020).
35. H. Wang, X. Du, W. Yu, Q. Chen, K. Zhu, Z. Chu, L. Yan and Y. Guan, Apollo's oracle: Retrieval-augmented reasoning in multi-agent debates (2023).
36. T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. V. Chawla, O. Wiest and X. Zhang, Large language model based multi-agents: A survey of progress and challenges, in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, ed. K. Larson (International Joint Conferences on Artificial Intelligence Organization, 8 2024). Survey Track.
37. Llama Team, AI @ Meta, The llama 3 herd of models `https://ai.meta.com/research/publications/the-llama-3-herd-of-models/`, Access July 28, 2024.
38. Y. Yu, C.-H. H. Yang, J. Kolehmainen, P. G. Shivakumar, Y. Gu, S. Ryu, R. Ren, Q. Luo, A. Gourav, I.-F. Chen, Y.-C. Liu, T. Dinh, A. Gandhe, D. Filimonov, S. Ghosh, A. Stolcke, A. Rastrow and I. Bulyko, Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition, *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* , 1 (2023).
39. B. Jimenez Gutierrez, N. McNeal, C. Washington, Y. Chen, L. Li, H. Sun and Y. Su, Thinking about gpt-3 in-context learning for biomedical ie? think again, in *Findings of the Association for Computational Linguistics: EMNLP 2022*, (Association for Computational Linguistics, 2022).
40. M. Moor, O. Banerjee, Z. Abad, H. Krumholz, J. Leskovec, E. Topol and P. Rajpurkar, Foundation models for generalist medical artificial intelligence, *Nature* **616** (2023).
41. D. Weissenbacher, X. Zhao, J. R. C. Priestley, K. M. Szigety, S. F. Schmidt, K. O'Connor, I. M. Campbell and G. Gonzalez-Hernandez, Biocreative viii – task 3: Genetic phenotype normalization from dysmorphology physical examinations (2023).
42. J. Wei, X. Wang, D. Schuurmans *et al.*, Chain-of-thought prompting elicits reasoning in large language models (2023).
43. Lightning AI, Litgpt `https://github.com/Lightning-AI/litgpt`, Accessed October 1, 2024.
44. G. Jones, J. Amoah, E. Y. Klein, H. Leeman, A. Smith, S. Levin, A. M. Milstone, K. Dzintars, S. E. Cosgrove and V. Fabre, Development of an Electronic Algorithm to Identify in Real Time Adults Hospitalized With Suspected Community-Acquired Pneumonia, *Open Forum Infectious Diseases* **8**, p. ofab291 (2021).
45. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes, *Annals of Internal Medicine* **156**, 11 (2012), PMID: 22213490.
46. A. A. Vazquez Deida, D. J. Bizune, C. Kim, J. M. Sahrmann, G. V. Sanchez, A. L. Hersh, A. M. Butler, L. A. Hicks and S. Kabbani, Opportunities to Improve Antibiotic Prescribing for Adults With Acute Sinusitis, United States, 2016–2020, *Open Forum Infectious Diseases* **11**, p. ofae420 (2024).
47. C. Peng, X. Yang, A. Chen *et al.*, A study of generative large language model for medical research and healthcare, *npj Digital Medicine* **6** (2023).
48. D. J. Livorsi, C. M. Linn, B. Alexander, B. H. Heintz, T. A. Tubbs and E. N. Perencevich, The value of electronically extracted data for auditing outpatient antimicrobial prescribing, *Infection Control 38; Hospital Epidemiology* **39**, p. 64–70 (2018).
49. T. Xiangru, Z. Anni, Z. Zhuosheng, L. Ziming, Z. Yilun, Z. Xingyao, C. Arman and G. Mark, Medagents: Large language models as collaborators for zero-shot medical reasoning (2024).