Clinical Panda: A Large Language Model for Diagnostic Explanation

Raj Krishnan Vijayaraj[a,*], **Shurui Zhou**[a] **and Mark Chignell**[b]

[a]Department of Electrical and Computer Engineering, University of Toronto, Canada
[b]Department of Mechanical and Industrial Engineering, University of Toronto, Canada

Abstract. Stringent privacy regulations restrict access to real-world patient data, hindering the development of advanced diagnostic tools. We introduce Clinical Panda, a Large Language Model (LLM) trained to generate evidence-based explanations for diagnoses from synthetic clinical notes. A curated dataset is generated from the question-answer pairs from DDXPlus to simulate realistic clinical scenarios. This approach allows practical training with a minimal dataset of just 1,000 samples, highlighting Clinical Panda's data efficiency. We also propose a novel input perturbation analysis technique that simulates missing information in clinical notes by systematically masking medical entities. This analysis reveals that Clinical Panda exhibits superior explanation quality and factual retention compared to established models, even with significant knowledge gaps. Clinical Panda retains 18.18 % more medical entities from the reference explanation than GPT 3.5 Turbo. Clinical Panda's ability to generate evidence-based explanations for diagnoses positions it as a promising tool for improving patient care, as supported by our findings on diagnostic explainability.

1 Introduction

Our research explores the intersection of advanced language models and clinical expertise. The application of Artificial Intelligence in healthcare applications is currently constrained by challenges posed by data scarcity and privacy concerns. Electronic health records (EHRs) contain valuable clinical information beyond structured data, but collecting and annotating this text requires significant resources and raises ethical considerations. These limitations hinder the development of deep learning approaches, including Large Language Models (LLMs).

While recent works like [25, 2, 15] demonstrate the potential of LLMs in healthcare, grounding them in factual knowledge remains a challenge. This raises concerns about their reliability for high-stakes medical applications requiring accurate results. This is why explanations generated by LLMs could play a key role in improving the reliability of AI systems in healthcare. Such an explanation would help medical professionals understand an AI model's reasoning behind a diagnosis, improving trust and enabling them to refine the diagnosis or treatment plan.

One of the recent works [1] has established the capacity of LLMs to generate clinical notes from patient encounters. Hence, LLMs can be employed to generate silver-standard clinical notes based on patient encounters, and these clinical notes could be used to build clinical capability for LLMs. The work [14] uses synthetic clinical notes to create clinical LLMs capable of doing multiple tasks. This points to a new direction of domain-adapting LLMs using synthetic data.

Our work focuses on building diagnostic explanation capability for LLM using synthetic clinical notes while maintaining data efficiency. Our research extends into the realm of explainability, a vital consideration across various classification methods in healthcare applications. Healthcare professionals develop diagnostic capabilities through extensive clinical experience and navigate complex patient cases by drawing on accumulated knowledge. Although the capacity to answer questions is valuable, our study focuses on the critical domain of explainability in medical LLMs.

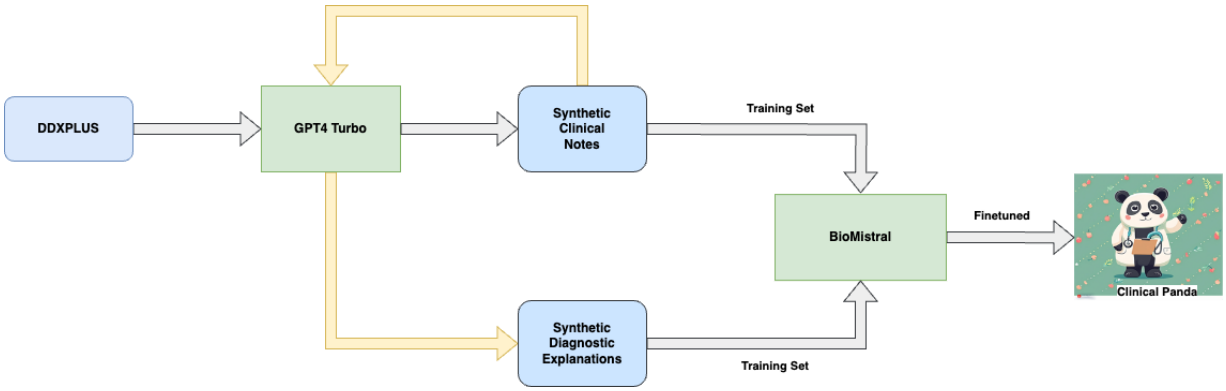
Explainable AI methods offer various techniques for generating explanations of classification decisions. Most methods like [18, 20, 23] can explain such decisions using a feature importance score. Within the framework of LLMs, the features correspond to word tokens, each receiving an assigned feature importance score. This can be used to highlight the most influential words from the input. However, in this form, these methods cannot articulate the underlying reasons in a human-readable manner. Hence, in this paper, we develop explanations in human-readable text format.

Our research aims to improve the explainability of medical diagnoses from clinical notes and provide practical solutions for healthcare professionals. We generate silver standard synthetic notes and explanations using the DDXPlus [24] dataset and use it to improve the diagnostic capability of medical LLM. By illustrating the rationale behind medical decisions, this study demonstrates how to evaluate explanations generated for diagnoses through LLMs.

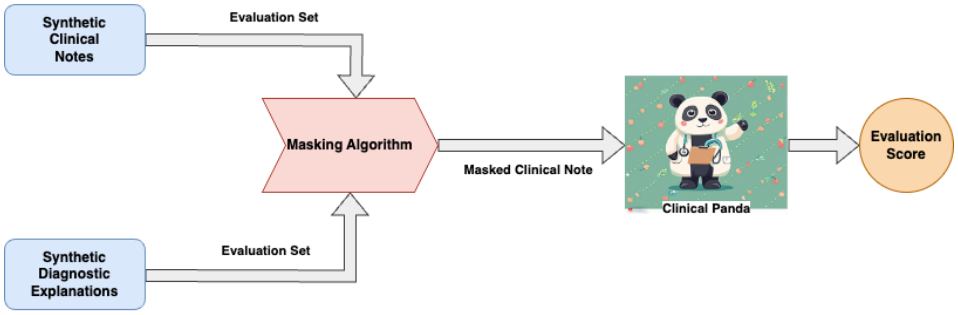
The following contributions are made in this paper.

- We introduce a novel framework that leverages LLMs to synthesize clinical notes from patient question-answer pairs. This approach directly addresses the challenge of data scarcity in medical applications by creating a training corpus for LLMs.
- We propose Clinical Panda, a Large Language Model designed to generate evidence-based explanations for diagnoses derived from clinical notes. Clinical Panda is trained using the synthetic clinical data generated by our framework.
- We propose a novel, lightweight perturbation analysis to assess the model's factual retention capabilities. This technique simulates missing information in clinical notes by systematically masking medical entities and evaluating model performance on masked data, comparing against the reference explanation using standard metrics

* Corresponding Author. Email: raj.vijayaraj@mail.utoronto.ca



(a)



(b)

Figure 1: Illustration of key stages in the development and evaluation of Clinical Panda: (a) Training and development phase showcasing the synthesis of clinical notes and model training; (b) Evaluation phase demonstrating the assessment of explanations generated under input perturbation.

2 Related Work

The recent development in open-source LLMs through works like [26] has led to the development of domain-specific LLMs. Clinical or medical LLMs like [2, 4, 25] excel at answering medical questions but require vast amounts of training data, a scarcity in real-world settings.

Numerous publicly available datasets provide valuable resources for training and evaluating models like Clinical Panda. MIMIC-III [11] is an example, while DDXPlus [24] presents a dataset tailored for automated diagnosis, featuring question-answer pairs and differential diagnoses. This gives space for expanding on developing synthetic clinical notes from question-answer pairs and related information.

Research in explainable AI aims to elucidate the decision-making processes of neural networks. Noteworthy contributions include [13], which evaluates explanation methods against selected explainability techniques, and the work [10], which assesses self-explanation capabilities against other explainable techniques. Additionally, works such as [27] explore innovative prompting strategies to enhance reasoning capabilities, while [9] summarizes current research in this area. Smaller fine-tuned models outperform LLMs in many biomedical reasoning or classification tasks.

The paper [16] showed that explanations from larger language models can increase the reasoning capabilities of smaller language models. Building on this concept [21, 19] distilled the reasoning abilities of larger Language Models into smaller, more efficient models. In other work [7], it was found that using more reasoning steps results in better reasoning results.

Causal reasoning is the analytical process of discerning the directional relationships between variables or events within a system, involving the identification of mechanisms, consideration of counterfactuals, and inference from observed evidence to establish cause-and-effect connections. [5] proposed a multi-step strategy for inference tasks using forward and backward chaining reasoning.

Similar to work [28], we used the DDXPlus dataset [24] to evaluate the diagnostic reasoning capability of LLMs. However, our work diverged from that previous research by prioritizing efficient training with synthetic data. We chose to use synthetic data to address the often-encountered real-world situations where stringent privacy regulations limit access to patient data.

Despite advancements in prompting LLMs for step-by-step reasoning and distilling reasoning abilities from larger models, current explainable AI methods for LLMs often yield unsatisfactory results. Feature importance scores are common but lack human-like reasoning and only highlight a few key factors. Additionally, smaller models can outperform LLMs in tasks that encourage reasoning. These limitations hinder the explainability of complex tasks, like medical diagnosis, where causal reasoning is crucial. We introduce Clinical Panda, a Large Language Model (LLM) trained to generate evidence-based explanations for diagnoses from synthetic clinical notes to address these limitations and overcome data scarcity in real-world clinical settings.

3 Clinical Notes and Explanations

The DDXPlus dataset [24] incorporates differential diagnosis, the consideration of multiple potential illnesses, and includes a broader variety of symptom and background information along with the question and answers for Automatic Diagnosis tasks.

We leveraged the answers to the questions and other diagnostic information to generate synthetic clinical notes. We used GPT-4 Turbo [22] to generate clinical notes and explanations for 1,100 records.

Clinical Notes. The study in [1] investigates how to generate clinical notes from patient-doctor encounters. A similar strategy using the DDXPlus dataset's [24] question-answer pair was followed in this work. Our process generated 1100 synthetic clinical notes, each containing information for 46 unique diagnoses/pathologies. The sections included in the synthetic clinical notes aim to capture a comprehensive picture of a patient's medical situation, similar to real-world clinical notes. Here is a breakdown of each section's purpose:

- **Patient Details**: This section establishes basic patient information like demographics and allergies, which are crucial for proper identification and medication management.
- **Chief Complaint**: This section captures the patient's primary reason for seeking medical attention, guiding the direction of the evaluation.
- **History of Present Illness** This section details the current illness's onset, progression, and associated symptoms, providing context for the current presentation.
- **Past Medical History**: This section documents past medical conditions, surgeries, and medications, offering insight into potential contributing factors.
- **Medications and Allergies** This section lists current medications and allergies, ensuring medication safety and avoiding potential interactions.
- **Physical Examination**: This section details the findings from a physical examination, providing objective data to support the diagnosis.
- **Assessment**: This section summarizes the physician's interpretation of the collected information, leading to a working diagnosis and explanation for the patient.
- **Plan**: This section outlines the proposed course of action, including treatment options, investigations, and follow-up recommendations.

Diagnostic Explanations. The explanations aim to provide a clear and concise understanding of the reasoning behind the diagnosis. The explanations also provide this reasoning with evidence from the given clinical notes. The GPT4 Turbo [22] was prompted with the diagnosis and the clinical note to generate an evidence-based diagnosis. The temperature parameter was standardized to 0 to utilize a greedy decoding strategy during text generation. The maximum number of new tokens was limited to 256 to limit the length of the explanation.

The fig. 1(a) clearly illustrates the generation of silver standard clinical notes from DDXPlus [24] dataset. The corresponding diagnostic explanation is also generated using the same pipeline. The generated clinical note and corresponding explanations are used to fine-tune the Mistral [31] into Clinical Panda.

Table 1: Token-based statistics of the evaluation set. The clinical note was generated with an upper limit of 512 tokens, and the explanation with an upper limit of 256 tokens. On average, explanations have a higher medical entity per token ratio.

Type	Tokens	Medical Entities
Clinical Note	453.38	29.05
Diagnostic Explanation	215.57	21.82

4 Clinical Panda

This section details the approach for fine-tuning an LLM to generate diagnostic explanations from synthetic clinical notes. We were inspired by the success of supervised fine-tuning with carefully curated training data, as demonstrated in the LIMA work. LIMA's [30] findings support the *Superficial Alignment Hypothesis*, suggesting that a relatively small set of high-quality examples can effectively guide an LLM towards accurate outputs, even without extensive human feedback or preference training. By leveraging this approach, we achieved efficient training for explanation generation.

To test this hypothesis within clinical diagnoses, we leveraged the dataset of synthetic clinical notes and explanations created as described earlier. These notes mimicked real-world patient records, incorporating details like demographics, presenting complaints, past medical history, medications, physical examination findings, and laboratory test results. To ensure the model evaluation is on unseen data, we randomly split the dataset into a training set of 1,000 samples and an evaluation set of 100 samples.

Clinical Panda architecture inherits the standard transformer architecture from Mistral [31]. We chose the Biomistral model[15] as the base model for fine-tuning. With approximately 7 billion parameters, Biomistral is one of Biomedical LLMs' most efficient open sources.

LoRA [8] based parameter-efficient fine-tuning technique was adopted to minimize computational resources during training. This approach leverages the advancements in Low-Rank Adaptation (LoRA) [8]. LoRA proposes injecting trainable low-rank matrices into specific model layers and training these layers alone, drastically reducing the number of trainable parameters compared to full fine-tuning.

QLoRA [6], an extension of LoRA, was leveraged for its memory efficiency and potential benefits. We employed QLoRA (Quantized Low-Rank Adaptation) to improve efficiency. By leveraging QLoRA, we loaded the pre-trained Biomistral model in a compressed format, significantly reducing its memory footprint and accelerating training on the V100 GPU. Additionally, QLoRA utilizes trainable Lower-Ranked Adapters (LRAs) instead of full-rank m trices in standard LoRA. These LRAs further reduced memory requirements and potentially improved the model's ability to learn from the synthetic clinical note data.

We fine-tuned the model on a single V100 GPU with a batch size 16. This batch size was chosen after careful experimentation to balance training efficiency with model convergence. Additionally, techniques like gradient clipping and learning rate scheduling were employed to optimize the training process and prevent overfitting.

5 Evaluation of Clinical Panda

Evaluating the quality of explanation generated by LLMs is an ongoing area of research. This work leverages three metrics commonly used for text summarization tasks: ROUGE score [17], Jaccard Similarity and BERTScore [29]. The ROUGE score measures the overlap of n-gram sequences (sequences of n words) between the generated explanation and generated references. We have used the precision value from the ROUGE metrics throughout. Jaccard similarity measures the overlap between key elements in two sets, which is useful for comparing the focus of explanations generated by different models. BERTScore, on the other hand, leverages pre-trained contextual embeddings and cosine similarity to assess the semantic similarity between explanations. We leverage the ROUGE score and BERTScore because prior research [1] has shown them well-

Algorithm 1 Perturbation Analysis to Evaluate Explanation

Require: $model$: Model to be evaluated Clinical Panda
$clinical\_note$: Clinical note as text
$reference\_explanation$: Silver standard reference explanation
ner_model : Medical named entity recognition (NER) model
Ensure: $explanation\_scores$: Dictionary containing lists of ROUGE and BERTScore for each masking level

$medical\_entities \leftarrow ner\_model(clinical\_note)$
$explanation \leftarrow model(clinical\_note)$
$baseline\_scores \leftarrow$ ComputeScores($explanation$, $reference\_explanation$)
$explanation\_scores \leftarrow \{\}$
$explanation\_scores["None"] \leftarrow baseline\_scores$
$masking\_step \leftarrow 0$
while $masking\_step \leq max\_masking\_level$ **do**
  $masked\_note \leftarrow$ MaskEntities($clinical\_note$, $medical\_entities$, $masking\_step$)
  $masked\_explanation \leftarrow model(masked\_note)$
  $masked\_scores \leftarrow$ ComputeScores($masked\_explanation$, $reference\_explanation$)
  $explanation\_scores[masking\_step] \leftarrow masked\_scores$
  $masking\_step \leftarrow masking\_step + 1$
end while

return $explanation\_scores$

correlated with human judgments of explanation quality. We have also used the Jaccard Similarity of medical entities to measure the medical relevance of the generated explanations.

We evaluated the diagnostic explanations generated by models using distinct evaluation metrics, namely Jaccard Score, BERTSCORE, ROUGE-1, and ROUGE-2. Each model's performance was assessed against these metrics to comprehensively understand their effectiveness in generating clinical dialogues. The correlation analyses presented in the results section highlight the varying degrees of alignment between the models across different evaluation metrics.

In addition to these established metrics, we introduce a novel perturbation analysis illustrated in Algorithm 1 to evaluate the factual retention of explanations. This simple and novel method uses masking of relevant medical entities from the clinical notes to impact the explanations generated by the model. We use a state-of-the-art medical named entity recognition (NER) model (MedCAT) [12] to identify these entities. The explanations are then re-generated, and these entities are masked progressively. The rationale behind this approach is that a great explanation should rely heavily on the core medical information, even if some of it is missing in the context. As relevant entities are masked, the quality of the explanation, as measured by ROUGE score and BERTScore, is expected to decrease significantly. This approach complements existing metrics by directly assessing the model's dependence on the factual content within the clinical notes.

To comprehensively evaluate Clinical Panda, we compared it against established LLMs: Mistral (domain-agnostic) and BioMistral (biomedical domain). Mistral serves as a general LLM baseline, while BioMistral offers a comparison within the biomedical domain. Additionally, GPT-3.5 Turbo, a powerful domain-agnostic LLM, was included. We randomly sampled 100 synthetic clinical notes (covering 33 diagnoses) and their explanations from a larger dataset (1100 notes). Importantly, the *Assessment* and *Plan* sections were excluded during evaluation, though included in the training. This ensures Clinical Panda focuses on explaining diagnoses, not treatment plans or physician assessments.

For the evaluation set, we generated explanations using Clinical Panda, BioMistral, and GPT-3.5 Turbo. Inference was performed on an A100 GPU for each available model (using API for GPT3.5 Turbo), with a batch size of 1 and consistent settings for temperature of 0, beam search parameters of 4, and maximum explanation length of 256 tokens. These parameters were carefully selected to standardize the explanation generation process across all models, ensuring a fair comparison.

We conducted the evaluation using unquantized models in all cases. This deliberate choice allows for a fair comparison of the models' inherent capabilities in explanation generation without introducing potential biases or performance variations due to quantization techniques.

The metrics to evaluate the impact of masking on explanation quality are the Jaccard Score, BERTSCORE, and ROUGE scores. ROUGE scores (ROUGE-1, ROUGE-2) assess the similarity between the generated explanation and the silver standard reference explanation by measuring the amount of overlap in sequences of words (n-grams). This approach effectively captures how well the model retains factual information from the clinical notes when faced with missing entities.

Jaccard Score provides a complementary perspective by directly assessing the overlap between the medical entities identified in the generated explanation and the silver standard reference explanation. This metric highlights the model's ability to focus on core medical concepts even when encountering masked elements in the input. To better understand how explanations degrade with increasing masking, we combined ROUGE scores, which focus on n-gram overlap, with the Jaccard Score, which emphasizes medical entity overlap. BERTScore leverages pre-trained contextual embeddings from the powerful BERT (Bidirectional Encoder Representations from Transformers) model. These embeddings capture the semantic relationships between words and phrases within a sentence. By comparing the BERTScore of the generated explanation with the silver standard reference explanation under increasing masking levels while Assessing how well the model preserves the semantic meaning of the explanation even when core medical entities are missing.

Table 2: Evaluation of the 1100 synthetic clinical notes generated by using DDXPlus dataset and GPT4 Turbo. These results show the explanation retains a subset of medical entities from the question-answer pairs.

Metrics	Value
Jaccard Score	0.26
Rouge-1	0.45
Rouge-2	0.17
Rouge-l	0.40

6 Results

This section presents results from experiments conducted on 100 random samples from the generated Clinical Notes and Explanations. The first sub-section below explores the characteristics of the synthetic clinical notes and explanations produced by the model. The next section focuses on the performance of Clinical Panda in generating diagnostic explanations. Its performance is compared against established LLMs using various metrics and a novel perturbation

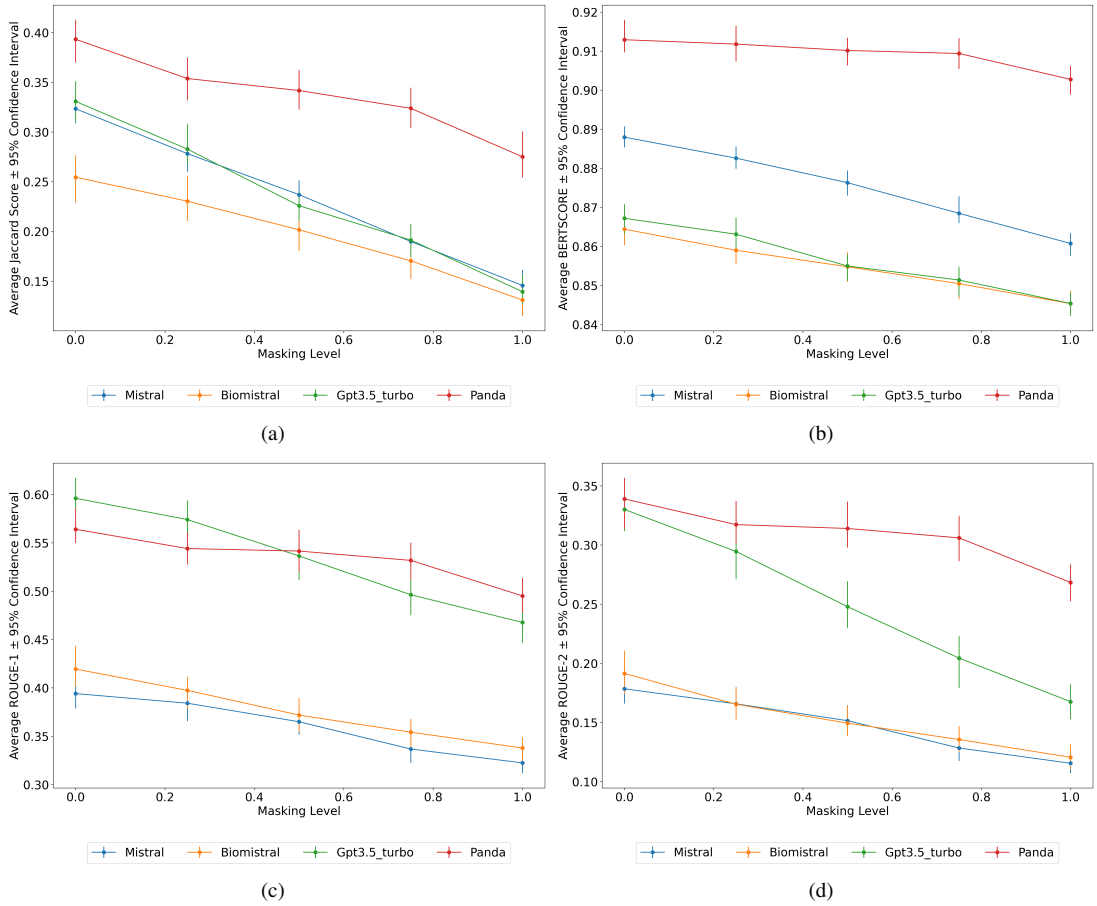


Figure 2: Performance of Clinical Panda and Compared LLMs under Increasing Masking Levels: This figure depicts the change in different evaluation metrics (a) Jaccard Score (medical entity overlap), (b) BERTSCORE (semantic similarity and fluency), and (c-d) ROUGE Scores's precision (overlap in word sequences) as the amount of masked information in the clinical notes increases (masking level). Error bars represent 95% confidence intervals obtained through bootstrapping. A decreasing trend can be observed throughout these graphs. This trend is the least for Clinical Panda, illustrating its diagnostic explanation capability even in the case of missing information.

analysis. Insights from perturbation analysis by masking medical entities are also discussed in this section and illustrated visually through Fig.2 and Table 3.

6.1 Clinical Notes and Explanations

Table 1 summarizes vital statistics about the generated data. This section reports the average number of tokens (words or sub-words) found in the clinical notes and explanations. This information provides insight into the overall length and complexity of the generated text. The table also reports the average number of medical entities identified within the notes and explanations. The average number of tokens in the clinical notes is below the threshold of 512, which was set during the generation phase. The average number of tokens in the diagnostic explanation is below the set value 256.

Table 2 summarizes the evaluation metrics used to assess the quality of the generated explanations. These metrics, such as the Jaccard Score, ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L), and BERTSCORE, measure different aspects of explanation quality. Jaccard Score and ROUGE scores focus on the overlap between the generated explanation and a reference explanation generated by GPT4 Turbo. BERTSCORE assesses the generated text's fluency and semantic similarity to reference text.

While quantitative metrics in Table 2 assess explanation quality, the provided example in Table 3 offers a qualitative view, completely understanding the model's strengths and weaknesses.

Qualitatively, the explanations are numbered, and evidence is pointed out from the clinical note. Often supported by additional medical information regarding that evidence. This initial non-expert analysis considers the semantics of the explanation, while medical relevance has to be established through expert evaluation.

6.2 Clinical Panda

This section evaluates the performance of Clinical Panda in generating diagnostic explanations for clinical notes.

We used the following metrics: Jaccard Similarity, BERTSCORE, and ROUGE scores (ROUGE-1, ROUGE-2). Jaccard Similarity measures the overlap between medical entities in the generated and silver standard reference explanations. Hence, Jaccard Similarity measures the medical relevance of the explanation. BERTSCORE evaluates the model's ability to capture semantic similarity. ROUGE scores assess the overlap between n-grams (sequences of words) in the generated and reference explanations. ROUGE scores are not domain-specific and give a general perspective on the similarity. It is also noteworthy to highlight that the Jaccard Similarity of medical entities is the best choice to measure the similarity grounded on the medical domain.

Table 3: This table shows the explanation for a case with URTI as a diagnosis for varied masking levels. Masking progresses 25% at a time. JS represents Jaccard Score and R-2 represents ROUGE-2 score.

Masking Level	Explanation	JS	R-2
0	The evidences for the diagnosis of an upper respiratory tract infection (URTI) in this patient are: 1. Symptomatology: The patient presents with a sore throat, increased sweating, and pain in various locations including the top of the head, left cheek, and left temple. The pain is described as tedious, sensitive, and heavy. The patient denies any radiation of the pain. The onset of the pain was sudden. The patient also reports having a fever, nasal congestion, a clear runny nose, and a cough. 2. Exposure History: The patient has been in contact with a person with similar symptoms in the past 2 weeks and works in a daycare. 3. Physical Examination: The patient appears alert and oriented. The patient is experiencing pain in the head and left side of the face. The patient also presents with a sore throat, fever, nasal congestion, and a cough. These clinical findings are consistent with the typical presentation of an upper respiratory tract infection.	0.52	0.35
1	The evidences for the diagnosis of an upper respiratory tract infection (URTI) in this patient are: 1. Symptomatology: The patient presents with a sore throat, increased sweating, fever, nasal congestion, a clear runny nose, and a cough, which are common symptoms associated with URTIs. 2. Pain Distribution: The patient experiences pain in the forehead, right cheek, and left temple, which can be indicative of sinus involvement, a type of URTI. 3. Exposure History: The patient has been in contact with a person with similar symptoms in the past 2 weeks and works in a daycare, which increases the risk of exposure to respiratory infections. 4. Living Conditions: The patient lives with 4 or more people, which can also increase the risk of transmission of respiratory infections. Given these clinical findings, the patient's presentation is consistent with a URTI, and further diagnostic testing is warranted to confirm the diagnosis.	0.39	0.31
2	The evidences for the diagnosis of viral pharyngitis in this patient include: 1. Sore throat: A common symptom of viral pharyngitis. 2. Increased sweating: Can be associated with fever and the body's response to infection. 3. Pain in various locations: May indicate a systemic viral infection. 4. Contact with a person with similar symptoms: Suggests a contagious viral etiology. 5. Working in a daycare: Increases the likelihood of exposure to viral pathogens due to close contact with children, who are common carriers of viral infections.	0.17	0.25
3	The evidences for the diagnosis of viral pharyngitis in this patient include: 1. Sore throat: A common symptom of viral pharyngitis. 2. Increased sweating: Can be associated with the body's response to infection. 3. Pain in various locations: May indicate a systemic viral infection. 4. Contact with a person with similar symptoms: Suggests a contagious viral etiology. 5. No significant past medical history: Suggests an acute condition rather than a chronic or recurrent issue.	0.1	0.20
4	The evidences for the diagnosis of viral pharyngitis in this patient include: 1. Symptoms: The patient presents with a cough, hemoptysis (coughing up blood), fever, and localized pain in the tonsils, cheeks, and temples, which are common symptoms of pharyngitis. 2. Exposure: The patient has been in contact with a person with similar symptoms in the past 2 weeks and attends a daycare, which increases the likelihood of exposure to viral pathogens. 3. Immunocompromised Status: As the patient is immunosuppressed, she is more susceptible to viral infections. 4. Absence of Travel: The lack of recent travel reduces the likelihood of exotic or travel-related infections. Given these clinical findings, a throat swab for rapid antigen detection of common viral pathogens such as influenza and adenovirus is warranted.	0.24	0.23

Table 4: Correlation Matrix for models using Jaccard Score. Clinical Panda correlates most to the Mistral response and the least with Biomistral.

	Panda	GPT3.5 Turbo	Biomistral	Mistral
Panda	1.000000	0.406640	0.338358	0.540308
GPT3.5 Turbo	0.406640	1.000000	0.415147	0.563479
Biomistral	0.338358	0.415147	1.000000	0.396297
Mistral	0.540308	0.563479	0.396297	1.000000

Table 5: Correlation Matrix for models using BERTSCORE. Clinical Panda correlates most to the GPT3.5 model response and the least with Biomistral.

	Panda	GPT3.5 Turbo	Biomistral	Mistral
Panda	1.000000	0.651927	0.512045	0.530808
GPT3.5 Turbo	0.651927	1.000000	0.443479	0.470723
Biomistral	0.512045	0.443479	1.000000	0.492231
Mistral	0.530808	0.470723	0.492231	1.000000

Table 6: Correlation Matrix for models using ROUGE-1. Clinical Panda correlates most to the GPT3.5 model response and the least with Biomistral. Correlation is also very strong with Mistral.

	Panda	GPT3.5 Turbo	Biomistral	Mistral
Panda	1.000000	0.363150	0.293800	0.357673
GPT3.5 Turbo	0.363150	1.000000	0.302239	0.290291
Biomistral	0.293800	0.302239	1.000000	0.315056
Mistral	0.357673	0.290291	0.315056	1.000000

Table 7: Correlation Matrix for models using ROUGE-2. Clinical Panda correlates most to the Mistral response and the least with Biomistral.

	Panda	GPT3.5 Turbo	Biomistral	Mistral
Panda	1.000000	0.303956	0.302373	0.402066
GPT3.5 Turbo	0.303956	1.000000	0.265286	0.205261
Biomistral	0.302373	0.265286	1.000000	0.457463
Mistral	0.402066	0.205261	0.457463	1.000000

Table 8: Evaluation of 100 explanations (with no masking) across models (JS is Jaccard Score and BS is BERTSCORE)

Model	JS	BS	R-1	R-2	R-L
Mistral	0.32	0.89	0.39	0.18	0.36
Bio-Mistral	0.25	0.86	0.42	0.19	0.39
GPT 3.5 Turbo	0.33	0.86	0.60	0.33	0.58
Clinical Panda	0.39	0.91	0.56	0.34	0.55

Table 9: Evaluation of 100 explanations (with masking level of 4) across models (JS is Jaccard Score and BS is BERTSCORE)

Model	JS	BS	R-1	R-2	R-L
Mistral	0.15	0.86	0.32	0.12	0.29
Bio-Mistral	0.13	0.85	0.34	0.12	0.31
GPT 3.5 Turbo	0.14	0.84	0.46	0.16	0.44
Clinical Panda	0.28	0.90	0.50	0.26	0.47

This is followed by ROUGE-2 as an n-gram of 2 would capture more relevant medical patterns in the text.

The correlation analyses of various models using different evaluation metrics reveal relationship patterns between the compared models. When assessing performance via Jaccard Score (Table 4), Clinical Panda exhibits the strongest correlation with Mistral responses while demonstrating the weakest association with Biomistral responses. Similarly, when evaluated using BERTSCORE (Table 5), Clinical Panda showcases the highest correlation with GPT3.5 Turbo responses, with Biomistral displaying the least alignment. ROUGE-1 assessments (Table 6) demonstrate Clinical Panda's strongest correlation with GPT3.5 Turbo. Moreover, in the context of ROUGE-2 (Table 7), Clinical Panda displays the highest correlation with Mistral, contrasting with Biomistral. This is the least expected result as Clinical Panda is fine-tuned on Biomistral [15]. It is possible that the 1000 samples used to fine-tune the model changed the response structure and the tokens frequently used to a very large extent. Further analysis is required to confirm this hypothesis. Clinical Panda's explanations showed the most coherence with either GPT-3.5 Turbo or Mistral, depending on the metric, while BioMistral's explanations diverged the most.

Table 3 showcases an example of a synthetic clinical note and its corresponding explanation generated by Clinical Panda for varying masking levels. This example focuses on a patient diagnosed with an Upper Respiratory Tract Infection (URTI). Analysis of the clinical note and explanation content and structure reveals that the numbered explanations reference specific evidence in the clinical note. It is also visible that the word count, Jaccard Score, and ROUGE-2 Score have a decreasing trend as the masking increases. This shows how the explanation becomes less informative and inaccurate as masking progresses.

Table 8 summarizes the performance of Clinical Panda compared to other models (Mistral, Bio-Mistral, GPT 3.5 Turbo) for the original clinical notes, without any masking applied. Clinical Panda achieved a Jaccard score of 0.39, indicating a good overlap of medical entities with the reference explanations. While BERTSCORE showed minimal variation across models, Clinical Panda maintained a high score of 0.91. ROUGE-1 revealed a more nuanced picture. GPT-3.5 Turbo achieved higher ROUGE-1 scores when the clinical notes were not masked. The higher ROUGE-1 scores achieved by GPT-3.5 Turbo in the unmasked condition might be caused by overfitting to the specific phrasing in the notes using techniques like Reinforcement Learning with Human Feedback [3]. Even then, the ROUGE Score of Clinical Panda is very close to that of GPT3.5 Turbo.

We conducted additional evaluations to assess how well the model handles missing information. We applied varying masking to the clinical notes in these evaluations, as shown in Fig.2. Under increased masking levels, Clinical Panda demonstrated superior performance across all metrics compared to other models. Table 9 presents the results obtained after masking all identifiable medical entities. This finding suggests that Clinical Panda is more robust regarding missing information and focuses on conveying the core meaning of the diagnosis.

In contrast, Mistral and Biomistral had a near-linear dependency on the masking level. Additionally, Fig. 2(a) highlights that GPT 3.5 Turbo produces significantly fewer medical entities in explanations with complete masking than Clinical Panda. A possible explanation for GPT3.5 Turbo showing this behaviour could be due to its immense training data; the model is giving out sound responses but with less focus on contextual medical information. This illustrates Clinical Panda's capacity to retain relevant medical information in explanations generated.

7 Conclusion

This study introduces Clinical Panda, a Large Language Model that enhances the explainability of diagnoses derived from clinical notes. We address the challenge of limited training data in the medical domain by generating synthetic clinical notes and diagnostic explanations using LLMs. These notes were generated from a question-answer dataset focused on patient signs, symptoms, and antecedents. This approach enabled Clinical Panda to be effectively trained with just 1,000 examples, achieving superior performance in three evaluation metrics compared to established models.

The correlation analysis revealed that Clinical Panda's explanations aligned more closely with those generated by Mistral and GPT-3.5 Turbo than BioMistral. We introduce a novel input perturbation analysis to assess Clinical Panda's robustness. This analysis simulated missing information (information gaps) and evaluated how well explanations held up under these conditions. The results demonstrated that Clinical Panda retained factual information more effectively than other models, highlighting its resilience even with limited data.

8 Limitations

While Clinical Panda demonstrates promise, further exploration is necessary. Evaluating its performance in a broader range of medical complexities and with a larger, more diverse dataset would provide more evidence of generalizability.

The next stage of development could incorporate human expertise. Medical professionals could analyze and refine Clinical Panda's explanations, ensuring alignment with best practices in clear and effective clinical communication. This human-in-the-loop approach can significantly enhance the quality, trustworthiness, and, ultimately, the patient experience with Clinical Panda's explanations. We propose incorporating expert evaluation in future work to strengthen our findings' reliability further and identify metrics that best align with expert judgment. Another direction is to explore the direction of RAG to augment the results with citations.

References

[1] A. Ben Abacha, W.-w. Yim, Y. Fan, and T. Lin. An empirical study of clinical note generation from doctor-patient encounters. In A. Vla-

chos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.168. URL https://aclanthology.org/2023.eacl-main.168.

[2] Z. Chen, A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, A. Sallinen, A. Sakhaeirad, V. Swamy, I. Krawczuk, D. Bayazit, A. Marmet, S. Montariol, M.-A. Hartley, M. Jaggi, and A. Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023.

[3] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

[4] C. Christophe, A. Gupta, N. Hayat, P. Kanithi, A. Al-Mahrooqi, P. Munjal, M. Pimentel, T. Raha, R. Rajan, and S. Khan. Med42 - a clinical large language model. 2023.

[5] A. Creswell, M. Shanahan, and I. Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3Pf3Wg6o-A4.

[6] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[7] Y. Fu, H. Peng, A. Sabharwal, P. Clark, and T. Khot. Complexity-based prompting for multi-step reasoning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yf1icZHC-l9.

[8] 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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[9] J. Huang and K. C.-C. Chang. Towards reasoning in large language models: A survey. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67.

[10] S. Huang, S. Mamidanna, S. Jangam, Y. Zhou, and L. H. Gilpin. Can large language models explain themselves? a study of llm-generated self-explanations, 2023.

[11] A. Johnson, T. Pollard, and R. Mark. MIMIC-III clinical database, 2023.

[12] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, and R. J. B. Dobson. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif. Intell. Med.*, 117:102083, July 2021. ISSN 0933-3657. doi: 10.1016/j.artmed.2021.102083.

[13] N. Kroeger, D. Ley, S. Krishna, C. Agarwal, and H. Lakkaraju. Are large language models post hoc explainers? In *XAI in Action: Past, Present, and Future Applications*, 2023. URL https://openreview.net/forum?id=mAzhEP9jPv.

[14] S. Kweon, J. Kim, J. Kim, S. Im, E. Cho, S. Bae, J. Oh, G. Lee, J. H. Moon, S. C. You, S. Baek, C. H. Han, Y. B. Jung, Y. Jo, and E. Choi. Publicly shareable clinical large language model built on synthetic clinical notes, 2023.

[15] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains, 2024.

[16] S. Li, J. Chen, Y. Shen, Z. Chen, X. Zhang, Z. Li, H. Wang, J. Qian, B. Peng, Y. Mao, W. Chen, and X. Yan. Explanations from large language models make small reasoners better, 2022.

[17] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

[18] S. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions, 2017.

[19] A. Mitra, L. Del Corro, S. Mahajan, A. Codas, C. Simoes Ribeiro, S. Agrawal, X. Chen, A. Razdaibiedina, E. Jones, K. Aggarwal, H. Palangi, G. Zheng, C. Rosset, H. Khanpour, and A. Awadallah. Orca-2: Teaching small language models how to reason. arXiv, November 2023. URL https://www.microsoft.com/en-us/research/publication/orca-2-teaching-small-language-models-how-to-reason/.

[20] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller. *Layer-Wise Relevance Propagation: An Overview*, pages 193–209. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_10. URL https://doi.org/10.1007/978-3-030-28954-6_10.

[21] S. S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, and A. Awadallah. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv: Computation and Language*, June 2023. URL https://www.microsoft.com/en-us/research/publication/orca-progressive-learning-from-complex-explanation-traces-of-gpt-4/.

[22] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct. 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

[24] A. F. Tchango, R. Goel, Z. Wen, J. Martel, and J. Ghosn. DDXPlus: A new dataset for automatic medical diagnosis. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=heBKnuV42O.

[25] A. Toma, P. R. Lawler, J. Ba, R. G. Krishnan, B. B. Rubin, and B. Wang. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding, 2023.

[26] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[27] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[28] C.-K. Wu, W.-L. Chen, and H.-H. Chen. Large language models perform diagnostic reasoning, 2023. URL https://openreview.net/forum?id=N0lQfjeNWOE.

[29] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with bert, 2020.

[30] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. YU, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=KBMOKmX2he.

[31] H. Zolkepli, A. Razak, K. Adha, and A. Nazhan. Large malaysian language model based on mistral for enhanced local language understanding, 2024.