Applied Intelligence (2026) 56:150
https://doi.org/10.1007/s10489-025-06827-1

Capability of large language models in assisting GPs with diagnoses

Ruibin Wang[1] · Abdul Rehman[1] · Tingting Li[1] · Rupert Page[2] · Hailing Li[3] · Xiaokun Wang[1,4] · Xiaosong Yang[1] · Jian Jun Zhang[1]

Accepted: 31 July 2025

Abstract

Purpose: A decision support pathway for general practitioners (GPs) was explored through automated referral letter analysis, with large language models' (LLMs) diagnostic roles comprehensively evaluated. **Methods:** The in-context learning performance of ChatGPT and GPT-4 for diagnostic decision support was evaluated using referral letters. Synthetic referral letters generated by ChatGPT addressed data scarcity, with distributional congruence quantified via Kullback-Leibler divergence. Two fine-tuning frameworks were comparatively assessed: encoder-based pre-trained language models (PLMs) for diagnostic classification, and decoder-based LLMs adapted to multiple-choice question-answering paradigms. **Results:** GPT-4 showed suboptimal few-shot accuracy (0.544). Synthetic letters demonstrated high fidelity (KL-divergence<0.05). Encoder-based PLMs consistently outperformed decoder-based LLMs when fine-tuned with augmented data, with BERT achieving 0.977 accuracy in mixed-train-collect-test protocols. Complementary F1 (0.9707) confirmed negligible diagnostic bias. **Conclusion:** LLMs exhibited insufficient diagnostic accuracy through both direct implementation (GPT-4 few-shot: 0.544) and fine-tuning approaches (accuracy 0.723), establishing fundamental limitations in clinical deployment. Crucially, their text-generation capability was leveraged for structured data augmentation, producing synthetic referral letters with high distributional fidelity (KL-divergence<0.05). This validated methodology enabled superior diagnostic performance through encoder-based PLM fine-tuning, where BERT achieved near-clinical-utility accuracy (0.977) - demonstrating 25.4% relative improvement over best-performing LLMs. Implementation pathways consequently prioritize this hybrid framework: LLM-mediated data augmentation followed by resource-efficient PLM classifiers, currently undergoing neurologist-piloted validation before multicenter expansion.

Keywords AI-based diagnosis · Referral letters · Data augmentation · Large language models · Clinical decision support

✉ Ruibin Wang
s5219587@bournemouth.ac.uk

Abdul Rehman
arehman@bournemouth.ac.uk

Tingting Li
tli@bournemouth.ac.uk

Rupert Page
rupert.page@uhd.nhs.uk

Hailing Li
lihailing@cuc.edu.cn

Xiaokun Wang
wangxiaokun@ustb.edu.cn

Xiaosong Yang
xyang@bournemouth.ac.uk

Jian Jun Zhang
jzhang@bournemouth.ac.uk

[1] National Centre for Computer Animation, Bournemouth University, Fern Barrow, Pool BH12 5BB, Dorset, UK

[2] University Hospitals Dorset NHS Foundation Trust, Longfleet Rd, Poole BH15 2JB, Dorset, UK

[3] Animation And Digital Art, Communication University of China, 1 Dingfuzhuang E St, Beijing 100024, China

[4] Institue of Artificial Intelligence, University of Science and Technology Beijing, No. 30 Xueyuan Road, Beijing 100083, China

Published online: 12 March 2026

1 Introduction

During the COVID-19 pandemic, England's National Health Service (NHS) faced considerable operational challenges. Secondary care capacity often struggled to accommodate GP-referred patients referred from GPs, leading to prolonged wait times and occasional referral denials . As the primary point of contact for patients, the efficiency of GPs is vital for expediting healthcare processes and ensuring prompt diagnosis and treatment. While GPs possess general knowledge, they lack specialization. An artificial intelligence (AI)-enabled assistant could enhance GP performance.

The digital transformation of healthcare has ushered in AI-assisted diagnostic platforms such as Symptomate[1], Patient[2], HealthTap[3], Ada[4], and other dialogue systems [1–4]. However, these systems are predominantly tailored for patients rather than GPs. While referral letters from GPs often encompass medical histories, clinical observations, current symptoms, and initial diagnoses, they also contain crucial information for diagnosis. However, the potential of these letters in assisting in disease diagnosis remains underexplored due to the challenges of labeling and anonymization. Henceforth, this study aims to employ the latest artificial intelligence techniques to extract features by training on referral letters, thereby assisting GPs in augmenting their diagnostic accuracy. To the best of our knowledge, this investigation represents the inaugural attempt to leverage referral letters for supplementary diagnostic purposes. In collaboration with the neurologists from the University Hospital Dorset (UHD), a dataset of labeled referral letters was assembled.

Due to the cost of labelling and anonymizing referral letters, the dataset size is limited. It is necessary to augment the data for downstream tasks. Existing methods include: Easy Data Augmentation (EDA) [5], BERT [6], BART [7], and CBERT [8]. These methods intensify the text's diverse representation. However, given that the representation and semantics of the text are constrained, the efficacy of these augmentation strategies may be limited. The advent of LLMs, notably ChatGPT and GPT-4 [9], has demonstrated profound capabilities across various domains. Drawing on the methodologies of data augmentation [10, 11], Alpaca [12] and ChatAug [13], this research harnesses the text generation prowess of ChatGPT to augment the assembled referral letters via prompt engineering, subsequently employing this dataset for downstream disease classification fine-tuning tasks.

The efficacy of ChatGPT and GPT-4 has been validated in common sense examinations like the USMLE [14] and medical question-answering systems such as MultiMedQA [15], where they exhibited remarkable performance [16]. Additionally, their competence in Name-Entity-Extraction (NER) tasks on the NCBI [17] and BC5CDR [18] datasets, as well as Relation Extraction (RE) tasks on GAD [19] and EU-ADR [20], has been evaluated, revealing unsatisfied results. Also, LLM platforms such as CancerGPT [21], Med-PaLM [15], and SynerGPT [22] have demonstrated potential in specialized medical domains. However, their efficacy in authentic clinical contexts for aiding GPs is yet to be established. Similarly, the potential of fine-tuning open-source LLMs such as LlaMA [23], Alpaca [12], GLM [24] for disease classification has not been investigated.

The primary goal of this study is to develop and validate a solution that supports GP in making accurate and professional diagnostic decisions using referral letters. In addition, the research aims to evaluate the effectiveness of LLM in aiding in disease diagnosis. The source code for this project is accessible at https://github.com/ruibin-wang/solution-assisting-GPs.git. The key contributions of this research can be summarized as follows.

- Evaluate the performance of LLMs, such as ChatGPT and GPT-4, in zero-shot, one-shot, and few-shot disease prediction tasks.
- Propose a data augmentation strategy using ChatGPT to overcome dataset limitations.
- Compare fine-tuning approaches for disease classification, including encoder-based PLMs and decoder-based LLMs.
- A recommended pathway involves utilizing ChatGPT for data augmentation, subsequently training encoder-based PLMs, and ultimately deploying the models to aid GPs in decision-making.

To the best of our knowledge, this is the first study to systematically explore the use of referral letters for AI-assisted GP decision support.

2 Method

The primary objective of this study is to find a solution that aids GPs in diagnostic decision-making accurately and professionally using referral letters, and to validate the efficacy of LLMs in disease diagnosis. To achieve this, we address three key research questions:

[1] https://digital.nhs.uk/services/e-referral-service/reports-and-statistics/appointment-slot-issue-reports

[2] https://symptomate.com/

[3] https://patient.info/

[4] https://ada.com/

- How effective are ChatGPT and GPT-4 in diagnosing diseases from referral letters using in-context learning?
- Can ChatGPT-generated synthetic data improve the performance of downstream disease classification tasks?
- Which fine-tuning strategy–encoder-based PLMs or decoder-based LLMs–performs better in classifying diseases from referral letters?

This section details our methodology, including: Data sources and preprocessing, in-context learning evaluation of ChatGPT and GPT-4, data augmentation using ChatGPT and fine-tuning strategies for disease classification.

2.1 Data sources

In collaboration with UHD Trust consultant neurologists, we collected a series of referral letters initially assessed by GPs and subsequently validated through specialist consultations - a two-stage diagnostic approach that mitigates potential bias from preliminary GP evaluations. Individual consent waived for this retrospective analysis of fully anonymized data. Following rigorous quality control, 439 unique cases representing 17 distinct neurological conditions were selected based on: (1) complete GP referral documentation; (2) specialist-confirmed neurological diagnoses. Exclusion criteria comprised referrals with insufficient clinical detail, non-neurological/inconclusive diagnoses, and duplicate episodes. Diagnostic labeling utilized ICD-10 codes (e.g., R56.9, M54.5, G40.3, G43.9) to ensure internationally standardized disease categorization and statistical comparability, with final codes assigned by consultant neurologists (multidisciplinary team consensus for complex cases). Patient privacy was protected through full anonymization (removal of names, addresses, NHS numbers, exact DOBs) and redaction of identifiable free-text details.

The developed methodology leverages structured GP referrals and specialist-validated ICD-10 labels with inherent scalability: while trained/validated on this dataset, the core framework can be adapted to larger, multi-institutional datasets upon demonstrated effectiveness. This future extension is central to enhancing model robustness and broad applicability across diverse clinical settings. A representative anonymized referral letter (Appendix A) illustrates GP assessments alongside neurologist-coded diagnoses, while Table 1 details the distribution of 17 neurological conditions.

Table 1 Case number for the 17 diseases

Disease name	Number
A41.9 - Sepsis, unspecified organism	16
G35 - Multiple sclerosis	34
G40.3 - Generalised idiopathic epilepsy and epileptic syndromes	18
G40.9 - Epilepsy, unspecified	20
G43.9 - Migraine, unspecified	35
G61.0 - Guillain-Barre syndrome	17
I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified	12
I63.9 - Cerebral infarction, unspecified	30
J18.9 - Pneumonia, unspecified organism	13
J69.0 - Pneumonitis due to inhalation of food and vomit	12
M54.5 - Low back pain	9
N39.0 - Urinary tract infection, site not specified	15
R29.8 - Other symptoms and signs involving the nervous and musculoskeletal systems	24
R41.0 - Disorientation, unspecified	11
R51 - Headache	50
R55 - Syncope and collapse	36
R56.9 - Unspecified convulsions	87

2.2 In-context learning for disease diagnosis using ChatGPT and GPT-4

The performance of ChatGPT and GPT-4, particularly in in-context learning and logical reasoning, is notably outstanding. This research employs two methodologies: direct diagnosis and a multiple-choice question-answering (QA) approach to evaluate their performance in disease diagnosis based on referral letters to assist GPs.

2.2.1 Direct diagnosis

This section evaluates the effectiveness of the GPT model in direct diagnosis using two prompt engineering approaches: zero-shot and one-shot methods. “Direct diagnosis" involves predicting the name of a disease or the ICD-10 code from a referral letter. Let $r_i \in R$ represent referral letters ($i = 1, \ldots, n$), and $d_j \in D$ denote the diagnosed diseases ($j = 1, \ldots, m$). A key instruction, *“Instruct = Please provide only the ICD-10 code or official disease name from the referral letter,"* was designed to ensure concise, accurate responses by limiting unnecessary output. This format supports clear comparison with actual diagnoses d_j and reflects clinical needs for precision. The prompt was used consistently as a system message before the input presentation. The predicted diagnosis for each r_i is denoted as p_i, and the model used–referred to as GPT–includes both ChatGPT and GPT-4. Details on input formulations are provided in Appendix B.

Zero-shot and one-shot approaches were employed to evaluate LLMs’ diagnostic accuracy without task-specific training or fine-tuning, motivated by their demonstrable general capabilities. Zero-shot evaluation specifically assesses the inherent diagnostic capability of models without examples, particularly relevant for novel clinical scenarios that lack training data. Simultaneously, one-shot evaluation examines the capacity for rapid adaptation from minimal demonstration data, a critical feature for practical deployment in resource-constrained clinical environments.

In the zero-shot process, the out-of-the-box capability of the GPT model is demonstrated through direct prediction of disease labels from referral letters without requiring annotated training data, which is defined as

$$p_i = GPT(Instruct, r_i). \tag{1}$$

A one-shot process occurs when a model learns to recognize a new category or task by using only one example. The one-shot process is defined as follows.

$$p_i = GPT(Instruct, example, r_i) \tag{2}$$

where the example is a pair consisting of an authentic referral letter and its corresponding diagnostic outcome, which can be found in Appendix B and is defined as:

$$example = [r_k, d_j]. \tag{3}$$

The prediction accuracy is used as a metric, which is defined as:

$$acc = \frac{sum(f(p_i, d_j))}{n}, \tag{4}$$

$$f(p_i, d_j) = \begin{cases} 1 & \text{if } p_i = d_j \\ 0 & else \end{cases} \tag{5}$$

Furthermore, based on previous findings that symptoms significantly influence the accuracy of the prediction in decoder-based PLM classification of referral letters [25], this study evaluates whether the association of extracted symptoms improves the diagnostic accuracy in ChatGPT and GPT-4. Symptoms s_k extracted from r_i using the IBM Clinical Annotator tool were integrated with original referral text using the "[SEP]" delimiter. This symptom-associated referral letter, denoted as r_{symp}, was employed to substitute r_i in the diagnostic process, structured as:

$$r_{symp} = r_i[SEP]s_1[SEP]s_2 \ldots [SEP]s_k \tag{6}$$

The integration aims to amplify key diagnostic signals and provide structured input for the GPT models, explicitly testing whether symptom augmentation enhances accuracy compared to using r_i alone. The input format is demonstrated in the "one-shot-symp" row of Appendix B, where extracted symptoms are highlighted in red.

2.2.2 Multiple-choice question and answering

To ensure reliable evaluation given the deterministic nature of disease diagnosis and the potential for uncontrolled, verbose outputs from generative models like ChatGPT and GPT-4, this paper complemented the direct diagnosis approach with a multiple-choice QA method. This constrained format presents the model with the referral letter, a clear diagnostic question, and a fixed set of possible disease options (including the true diagnosis d_j and $A - Q$ common distractors). The model is instructed to output only the corresponding option letter (e.g., options A, B, C,...,Q). This approach forces the model to focus solely on selecting the correct diagnosis from the provided choices and guarantees unambiguous, easily evaluable outputs, significantly simplifying and robustifying the accuracy assessment compared to parsing open-ended generation. (Example prompts in Appendix C).

The zero-shot process:

$$opt_i = GPT(Definition, r_i, [question, option_{set}]) \tag{7}$$

The one/few-shot process:

$$\begin{aligned} opt_i = GPT(&Definition, \\ &[r_k, question, option_{set}, d_j]_{few}, \\ &[r_i, question, option_{set}]) \end{aligned} \tag{8}$$

Where:

- Definition is phased as "For this task, we request that you make a diagnosis by examining the provided referral letter."
- $option_{set}$ represents the set of disease name options.
- opt_i denotes the predicted choice within $option_{set}$.
- The variables GPT, r_i, and d_j are consistent with the parameters defined in the direct diagnosis method.
- *question* serves as the directive for the model's output.

For illustrative purposes, an example of a multiple-choice QA prompt is provided in Appendix C. It is noteworthy that, for the few-shot process, five samples were chosen to help the GPT model understand the paradigm of the prediction task. And to make Appendix C more clear, the samples associated with the one-shot and few-shot processes have been designated with the colour blue. This deliberate distinction serves to differentiate these samples from the reference letter employed for prediction purposes.

2.3 Data augmentation

As highlighted in Section 1, the labor-intensive and expensive nature of labeling and anonymizing referral letters presents significant challenges, resulting in limited availability of public clinical text datasets. Through collaboration with neurologists at UHD, a labeled referral letter dataset was compiled (Table 1). However, the dataset size was considered insufficient for downstream neural network training, and an imbalanced distribution of disease cases was observed,

potentially compromising efficacy for subsequent tasks. Inspired by methodologies described in ChatAug [13], Alpaca [12], and ZeroShotDataAug [26], a dataset augmentation approach was implemented leveraging ChatGPT's prompted text generation capabilities, which demonstrated performance improvements in downstream disease classification tasks.

Augmentation methodology Let D denote disease names and R represent referral letters. Few-shot text generation capabilities of ChatGPT were employed to produce an augmented dataset R_{aug} (Algorithm 1). Two parameters controlled the process: $data_{num}$ (total instances generated) and $perturn_{num}$ (instances generated per API call). Due to observed latency in ChatGPT's API responses, $perturn_{num} = 2$ was empirically determined to optimize throughput. The procedure is illustrated in Fig. 1.a, where the prompt module corresponds to Algorithm 1.

API implementation details As shown in Algorithm 1, the prompt construction and model invocation procedure entailed concatenating three core components: a predefined instruction template *Instruct*, the target disease designation $disease_{name}$, and two clinically authentic referral letter exemplars. Specifically, the instruction string was programmatically merged with the disease identifier, followed by the sequential insertion of the reference letters labeled as "Referral letter 0" and "Referral letter 1". This composite prompt structure, integrating pedagogical guidance with domain-specific exemplars, was then submitted to ChatGPT's inference API. API responses gpt_{resp}, generated through this carefully engineered prompt,were parsed to extract generated text while stripping conversational artifacts.

Limitations and mitigation Important limitations associated with synthetic text generation warrant consideration, including potential propagation of biases present in base examples or model training data, possible emergence of repetitive syntactic patterns across generated samples, and risks of overfitting when downstream models train extensively on augmented data. To mitigate these concerns, maximum variation prompting using diverse anchor samples was implemented, generated samples underwent distributional similarity checks against original data, and downstream model regularization techniques such as dropout and weight decay were employed during fine-tuning.

Algorithm 1 The framework of using ChatGPT to augment referral letter dataset.

Input: Disease name set D, Referral letter R which contains the paired referral letter R_{referral} and its corresponding disease name R_{disease}. $R = \{R^i\} = \{[R^i_{\text{referral}}, R^i_{\text{disease}}]\}$ where $i = 1, \ldots, m$ represents the number of diseases

Initialization: Initialize chatgpt-3.5 API key, the generated dataset $R_{\text{aug}} = []$.

Parameters: $data_{num}$ is the total number of generated referral letters, $perturn_{num}$ is the number of generated referral letters per-turn. $Instruct =$ "You are a medical text generator, based on the following examples, please generate " + str($perturn_{num}$) + " referral letters for patients diagnosed with the disease "

for each $disease_{name}$ in D **do**
  for $turn$ in 0 to $data_{num}/perturn_{num}$ **do**
    $rd_{\text{referral}} = \text{random}(R_{\text{referral}}, 2)$ when $disease_{name} = R_{disease}$
    $prompt = Instruct + disease_{name}$ + "Referral letter 0: " + $rd_{\text{referral}}[0]$ + "Referral letter 1" + $rd_{\text{referral}}[1]$
    $gpt_{resp} = \text{ChatGPT}(prompt)$
    $R_{\text{aug}} = R_{\text{aug}} \cup [gpt_{resp}, disease_{name}]$
  end for
end for

2.4 Embedding visualisation and evaluation

This section presents a visualisation of both the augmented dataset and the originally collected dataset within a shared

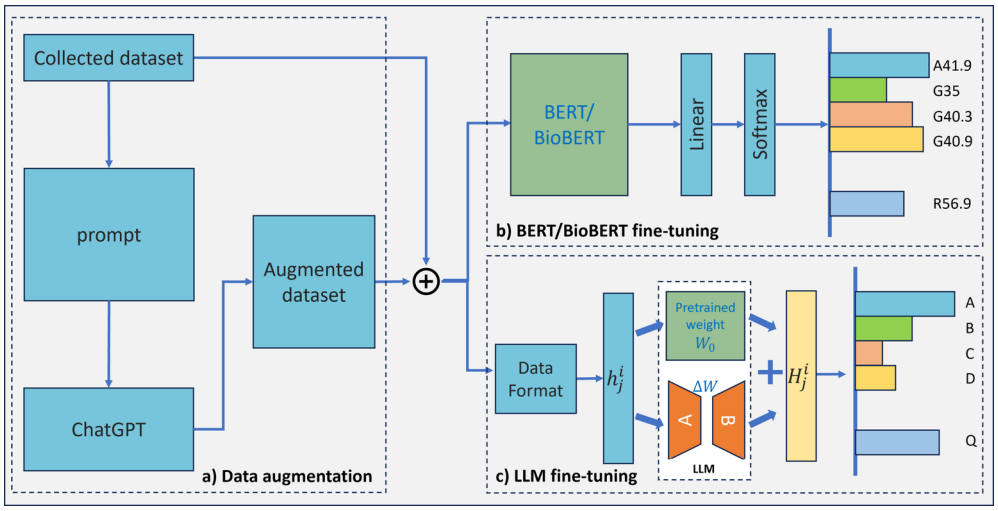

Fig. 1 Flowchart of data augmentation and downstream fine-tuning

embedding space, with the objective of determining the distributional proximity between synthetically generated and authentic referral letters. For embedding the referral letters, the BioBERT [27] model is employed, specifically selected for its domain-specific pre-training on extensive biomedical literature including PubMed abstracts and PubMed Central full-text articles. This specialized architecture demonstrates enhanced capability in capturing clinical semantics and ontological relationships inherent in medical documentation, thereby providing clinically relevant embeddings. The embedding $\phi(r)$ of a given referral letter r is formally defined as:

$$\phi(r) = \text{BioBERT}(r) \tag{9}$$

where $\phi(r)$ corresponds to the contextual representation of the [*CLS*] token extracted from the final transformer layer.

Prior to assessing the similarity between a generated referral letter and its corresponding actual referral letter for a specific disease, a formal definition is introduced. Analogous to the description in Section 2.3, the dataset of collected referral letters is denoted as $R = \{R^i\} = \{[R^i_{referral}, R^i_{disease}]\}$ where $i = 1, \ldots, m$ represents the number of diseases. For each $R^i_{referral} = \{r^i_j\}$ where $j = 1 \ldots n_i$ denotes the number of referral letters corresponding to disease i. Similarly, the augmented dataset is represented as $R_{aug} = \{R^i_{aug}\}$ for $i = 1, \ldots, m$, and for each $R^i_{aug} = \{r^i_{aug(j)}\}$ for $j = 1, \ldots, n^i_{aug}$ corresponds to the number of augmented referral letters for disease i. To quantitatively evaluate distributional alignment, Kullback-Leibler (KL) divergence is adopted [28]. This information-theoretic metric measures the information loss when approximating the true distribution $P_{\text{real}}(r|d_i)$ of disease-specific referral letters using the synthetic distribution $P_{\text{gen}}(r_{\text{aug}}|d_i)$ [29]. Formally:

$$D_{KL}(P_{\text{gen}} \parallel P_{\text{real}}) = \sum_{x \in \mathcal{X}} P_{\text{gen}}(x) \log \frac{P_{\text{gen}}(x)}{P_{\text{real}}(x)} \tag{10}$$

where $\mathcal{X}$ denotes the embedding space. Reduced KL values indicate enhanced distributional congruence, signifying that synthetic letters preserve statistical properties of authentic medical documentation. This metric provides quality verification at the distribution level, which is fundamental to ensuring clinical validity in downstream applications. Algorithm 2 implements the empirical computation, in which the resultant S^i_j quantifies the divergence per sample, with values asymptotically approaching zero confirming that synthetic referral letters reside within the characteristic manifold of genuine medical text while preserving clinically significant correlations.

Algorithm 2 Distributional Similarity Measurement via KL Divergence.

Require: Real dataset $R^i = \{r^i_1, \ldots, r^i_{n_i}\}$, Synthetic dataset $R^i_{\text{aug}} = \{r^i_{\text{aug}(1)}, \ldots, r^i_{\text{aug}(k)}\}$
Ensure: Similarity vector $S^i \in \mathbb{R}^k$
1: Initialize BioBERT encoder $\phi(\cdot)$ with clinical-domain weights [27]
2: **for** each synthetic sample $r^i_{\text{aug}(j)} \in R^i_{\text{aug}}$ **do**
3: Compute empirical divergence metric:

$$S^i_j = \frac{1}{|R^i|} \sum_{r^i_k \in R^i} \phi(r^i_{\text{aug}(j)}) \left(\log \phi(r^i_{\text{aug}(j)}) - \log \phi(r^i_k) \right)$$

▷ Monte Carlo approximation of $D_{KL}(P_{\text{gen}} \parallel P_{\text{real}})$
4: **end for**

2.5 Disease classification task

To identify the optimal solution for aiding GPs with referral letters, this section introduces several methods leveraging the augmented dataset. Concurrently, to assess the efficacy of fine-tuning LLMs on free-text-based disease classification, two distinct mechanisms are proposed: the encoder-based PLMs fine-tuning and the LLMs fine-tuning. The flowchart of the two mechanisms can be found in Fig. 1.b and .c.

2.5.1 Encoder-based PLMs Fine-tuning

Leveraging encoder-based PLMs' (BERT [6] and BioBERT) prominence in text embeddings, this paper train it with the augmented data, targeting at interpreting referral letters to support GP's decision making. Encoder-based PLMs are primarily pre-trained to minimize the cross-entropy loss:

$$L(d, p) = -\sum_{i=1}^{m} \sum_{j=1}^{n^i_{\text{aug}}} d^i_j \cdot log(p^i_j) \tag{11}$$

Tokens integrate both token and positional embeddings:

$$h^i_j = r^i_{\text{aug}(j)} W_e + W_p \tag{12}$$

where W_e is the token embedding matrix, W_p is the position embedding matrix and h^i_j is the tokenized vector for the input $r^i_{aug(j)}$. The encoder-based PLMs, namely BERT and BioBERT, are then employed to extract the embedding features from the input referral letters:

$$H_j^i = \text{encoder}_{model}(h_j^i) \tag{13}$$

Finally, a linear classifier with Softmax activation computes the disease distribution for input $r^i_{aug(j)}$:

$$p_j^i = \text{Softmax}(\text{Linear}(H_j^i)) \tag{14}$$

2.5.2 LLM fine-tuning

Before LLM fine-tuning, a data format module (as shown in Fig. 1.c) that reformats the augmented dataset into a multiple-choice QA format was employed. In this section, disease prediction categories are input choices, aiming to predict the right option. Each $r^i_{aug(j)}$ in R_{aug} consists of K tokens: $r^i_{aug(j)} = \{s_1, s_2, \ldots, s_K\}$. Fine-tuning the LLM seeks to maximize:

$$L\left(r^i_{\text{aug(j)}}\right) = \sum_{k=1}^{K} \log P\left(s_k \mid s_1, s_2, \ldots, s_{k-1}; \theta\right) \tag{15}$$

where θ denotes the trainable parameters of the LLM. Tokens are then represented as:

$$h_j^i = r^i_{\text{aug(j)}} W_e + W_p \tag{16}$$

with W_e and W_p defined as the same in the "Encoder-based PLMs Fine-tuning" section. The subsequent transformation is:

$$H_j^i = \text{LLM}(h_j^i) \tag{17}$$

As low-rank adaptation (LoRA) [30] and parameter-efficient fine tuning (PEFT) [31] techniques significantly reduce computational and storage requirements by fine-tuning only a minimal subset of parameters (typically $< 1\%$ of the model), enabling efficient adaptation even on hardware with limited resources. Crucially, these methods retain the integrity of pre-trained knowledge to prevent catastrophic forgetting while facilitating streamlined adaptation to multiple downstream tasks through modular weight updates or task-specific adapters. Therefore, this paper adopts PEFT supported LoRA to efficiently fine-tune the LLMs on referral letters. Considering computational constraints, this study utilized LlaMa-7B and LlaMa-13B for disease classification. As depicted in Fig. 1.c, for a pre-trained model's parameter matrix $W_0 \in \mathbb{R}^{d\times k}$, $\Delta W \in \mathbb{R}^{d\times k}$ denotes the fine-tuned parameters. Given ΔW's low dimension, it is expressed as:

$$W_0 + \Delta W = W_0 + BA \tag{18}$$

Here, B and A are low-rank trainable matrices, with $\text{rank}(A), \text{rank}(B) \ll \min(d, k)$, improving efficiency and memory usage. The forward pass $H_j^i = W_0 h_j^i$ in the LLM becomes:

$$H_j^i = W_0 h_j^i + \Delta W h_j^i = W_0 h_j^i + BA h_j^i \tag{19}$$

The target token prediction uses the linear matrix W_e^T as:

$$s_i = \text{Softmax}(H_j^i W_e^T) \tag{20}$$

3 Experiments and results

This section presents the experimental setups based on the methods from Section 2. Both direct and multiple-choice QA methods for disease diagnosis were evaluated using ChatGPT and GPT-4 in conjunction with referral letters. The data augmentation is explored using ChatGPT and analyzed the embeddings of the augmented versus original referral letters. The fine-tuning process for encoder-based PLMs and LLMs using our augmented dataset is also presented.

3.1 Hardware configuration

The experiments were conducted on a system equipped with Ubuntu 20.04.4 LTS, 128GB of system RAM, and dual NVIDIA RTX A5000 GPUs, each providing 24GB of usable memory. The system's CPU, an AMD Threadripper 3975WX, features 32 cores and 64 threads. This hardware configuration, particularly the dual-GPU setup, was selected to meet the significant memory demands inherent in processing Llama-7B model. Loading this model alone for inference, even under a parameter-efficient fine-tuning approach Like LoRA, requires at least 24GB of GPU memory per instance, exclusive of the additional overhead associated with processing input data embeddings. Given the aggregate memory requirement exceeding 24GB per GPU instance and considering the available hardware resources within our laboratory, employing dual RTX A5000 GPUs offered the optimal balance of sufficient memory capacity (48GB aggregate) and computational capability for this large-scale model experimentation.

3.2 Baselines methods

This paper utilized multiple models for disease diagnosis via free-text referral letters from GPs.

ChatGPT and GPT-4 LLMs developed by OpenAI, accessible exclusively via API.

BERT BERT is the encoder module of the transformer and is pre-trained on a massive text corpus. Specifically, the "bert-base-cased" model was adopted, featuring 12 layers, 768 units, 12 attention heads, and a total of 110 million parameters.

BioBERT A BERT adaptation for biomedical texts, BioBERT is pretrained on vast biomedical corpora. It often outperforms standard BERT in biomedical tasks due to its domain-specific expertise. This paper chose "biobert-base-cased-v1.2", with the same specifications to BERT.

LlaMa Presented by Meta AI, LlaMa is trained on 1 trillion tokens derived from open-access sources. The selection of LlaMa-7B and LlaMa-13B was influenced by the available hardware specifications. It is worth mentioning that, due to the restrictions on hardware and budget limitations, the experiment with LlaMa-13B was only evaluated with the 500-case dataset.

3.3 Experiment design and results

The aims of this paper are to assist GPs in decision-making with referral letters and validate LLMs for disease classification. We used in-context learning diagnostics for ChatGPT and GPT-4, based on 439 collected referral letters, applied data augmentation, and fine-tuned PLMs. Detailed experiments and results are in the following section.

3.3.1 In-context learning for disease diagnosis using ChatGPT and GPT-4

Direct diagnosis on ChatGPT and GPT-4 *The diagnostic capabilities of ChatGPT and GPT-4 were evaluated using clinical referral letters with zero-shot and one-shot prompts based on Appendix A. Symptom information was incorporated into the prompts to assess its impact. Diagnostic accuracy with in-context learning is shown in Table 2, where zero-shot-symp and one-shot-symp columns indicate accuracy with symptom-inclusive prompts.*
Analysis of Table 2 yields the following insights:

1) One-shot accuracy exhibited a 21.5% increase over zero-shot for ChatGPT, compared to 9.1% for GPT-4, indicating sample integration enhances performance.
2) By integrating symptoms into the prompt, zero-shot accuracy increased by 0.2% for ChatGPT and 2.5% for GPT-4. In the one-shot phase, ChatGPT's accuracy rose by 1.1%, whereas GPT-4's increased by 0.2%. Integrating symptoms into the prompt enhances prediction accuracy, albeit marginally.

Table 2 The accuracy of direct diagnosis on ChatGPT and GPT-4

Model	zero-shot	zero-shot-symp	one-shot	one-shot-symp
ChatGPT	0.118	0.120	0.333	0.344
GPT-4	0.203	0.228	0.294	0.296

Table 3 Accuracy of multiple-choice QA disease prediction

Model	zero-shot	one-shot	Few-shot
ChatGPT	0.141	0.355	0.405
GPT-4	0.100	0.394	0.544

Multiple-choice question and answering on ChatGPT and GPT-4 *In this section, the GP disease diagnosis was restructured into a multiple-choice QA format, using disease categories as answer options. Following the descriptions in the Section 2, the prompt from Appendix C was adopted, and the results are presented in Table 3.*

Table 3, compared to direct diagnosis in Table 2, shows a 2.3% improvement in zero-shot prompt precision for ChatGPT and a 10.3% decrease for GPT-4. The prompt accuracy for one-shot increases by 2. 2% for ChatGPT and 10% for GPT-4. In addition, Table 3 highlights the enhanced prediction accuracy of the few-shot prompt for both models.

In particular, even the highest accuracy achieved (GPT-4: 0.544) is still not reliable for clinical use. In practice, a precision of more than 90% is generally required for a diagnostic tool to be considered useful. This limitation is due to inconsistent reasoning and poor understanding of disease relationships in general-purpose LLMs like ChatGPT and GPT-4. Further experiments show that closing these diagnostic gaps requires fine-tuning or pre-training with specialized healthcare data, which can lead to clinically acceptable performance.

3.3.2 Data augmentation

To augment the dataset, Algorithm 1 was utilized, generating datasets of 100, 200, and 500 cases per disease. These subsets were subsequently amalgamated into a composite 800-case dataset. The BioBERT encoder was employed to derive embedding representations for each referral letter. A visualization of both collected and generated referral letters within the same embedding space is presented in Fig. 2. In this representation, blue points denote embeddings derived from actual clinical referral letters, while orange points signify those from the augmented data. Although several orange points are distributed peripherally, a predominant central clustering is observed, indicating a high degree of representational

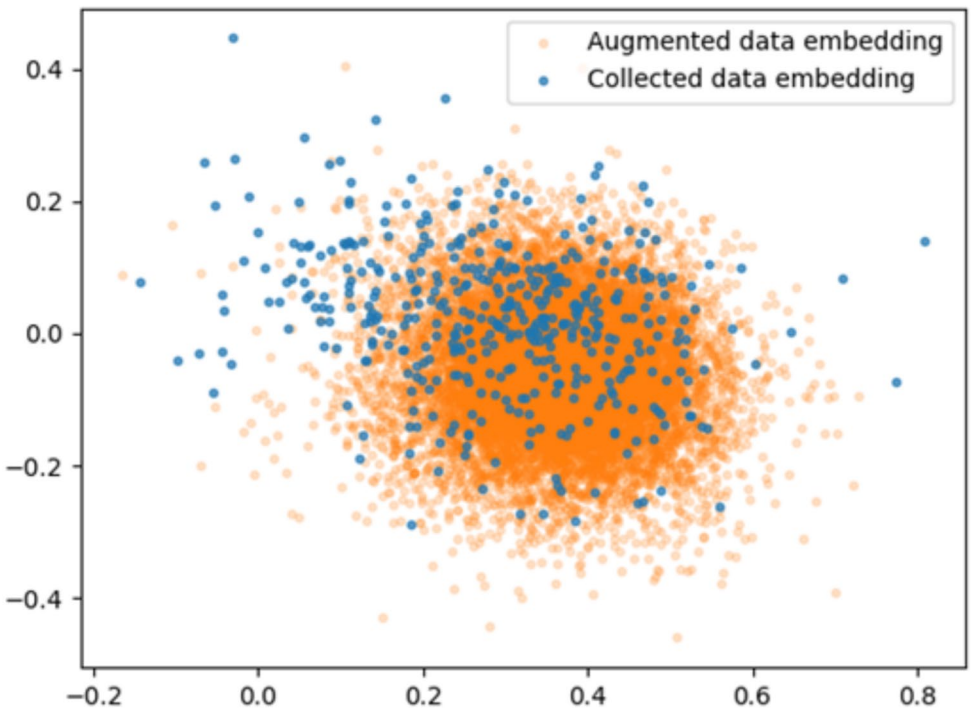

Fig. 2 Visualization of referral letters in the embedding space

congruence. This spatial alignment corroborates the suitability of the generated data for augmentation purposes.

Further quantitative validation was performed using Algorithm 2, which assessed the similarity between generated and collected referral letters. The resulting KL divergence values, depicted as a boxplot in Fig. 3, demonstrate divergence values predominantly below 0.05 for most diseases, providing statistical evidence for the semantic proximity between generated and authentic letters. This high degree of similarity is anticipated to enhance the learning efficacy in subsequent data-intensive deep learning applications. Additionally, the presence of outliers for each disease, as evident in the boxplot, introduces valuable distributional variation that can positively contribute to model generalizability. However, tight concentration near the 0.05 threshold suggests potential latent overfitting risks in downstream tasks, necessitating deliberate regularization strategies during fine-tuning. This risk mitigation strategy was integrated into the subsequent experimental design detailed in Section 3.3.3 ("Fine-tuning on the PLMs and LLMs") through dropout optimization and validation metrics monitoring. The results shown in Fig. 4 confirm that no significant overfitting occurred.

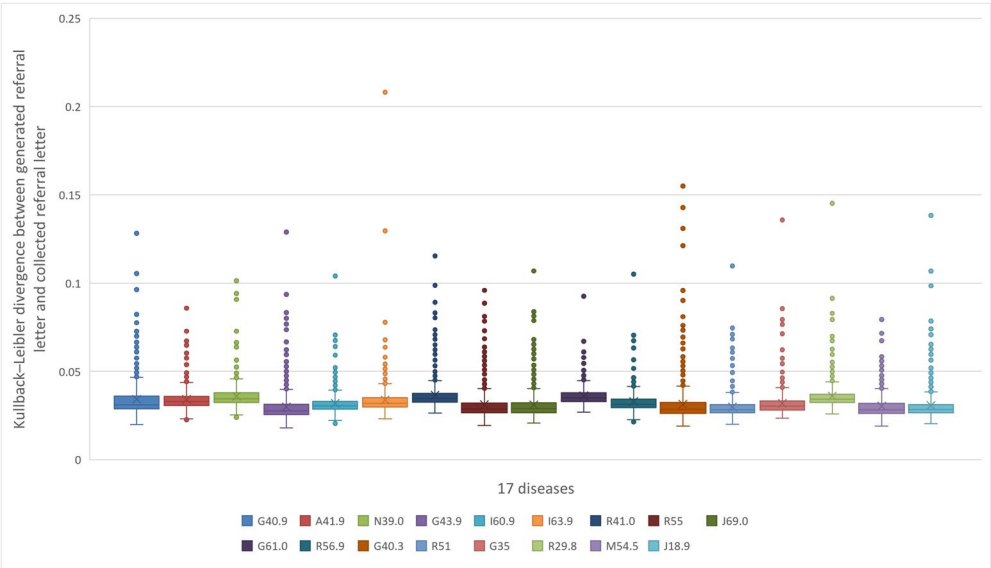

Fig. 3 Kullback–Leibler divergence between generated referral letters and collected referral letters

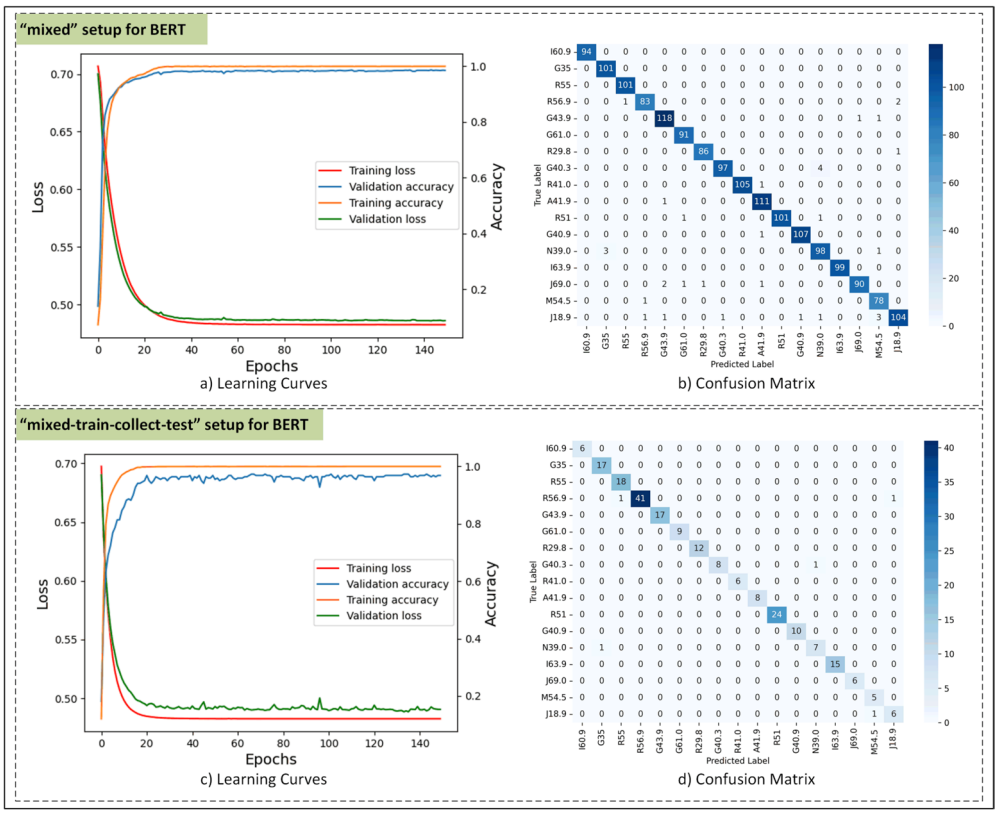

Fig. 4 Learning curves and confusion matrix of "mixed" and "mixed-train-collect-test" setups for BERT

3.3.3 Fine-tuning on the PLMs and LLMs

In this section, experiments were conducted on: 1) supervised fine-tuning using BERT and BioBERT, and 2) multiple-choice QA predictions using the LLMs. For PLMs (BERT/BioBERT), the model was trained for 100 epochs using the Adam optimizer with a learning rate of 1e-6 and a batch size of 8, with dropout set to 0.5. The dataset was split in a 3:1:1 ratio for training, test, and validation, respectively, translating to approximately 60% of the data for training and the remaining 40% equally divided between test and validation in the 100-case experiment. In contrast, LLMs (LlaMa-7B/13B) used the Adam optimizer with a higher learning rate of 1e-4, a batch size of 1, and leveraged LoRA adaptation (rank r=120), resulting in 92.5% trainable parameters. Their data set used a 3:1 train / test split, where the size of the test set was scaled with the volume of the data (for example, approximately 26% for the 100-case experiments), without reference to an explicit validation set. The effectiveness of both augmented and collected data in downstream fine-tuning tasks was assessed, and potential overfitting from a concentrated augmentation dataset was explored through four experimental setups:

- mixed: This setup combines both the generated referral letters and the actual collected referral letters. Both data types are randomly distributed across the training, testing, and evaluation datasets. This design seeks to assess model performance on the "mixed dataset" holistically, setting aside concerns of overfitting.
- gen-train-collect-test: This approach utilises the entire generated dataset for training the model, while the real collected dataset serves the purpose of evaluation. This paradigm specifically addresses the risk of data leakage by ensuring the test set remains uncontaminated by synthetic data, preventing knowledge boundary overlap within the test distribution. The objective is to determine the influence of the augmented dataset on downstream fine-tuning outcomes for both encoder- and decoder-based PLMs.
- mixed-train-collect-test: In this configuration, the training set consists of both generated and collected data, while the testing set exclusively uses real collected data. While it shares some similarities with the "gen-train-collect-test" paradigm in its efforts to circumvent overfitting, it integrates real collected data into the training set to bolster authentic knowledge transfer to the model. This could potentially enhance accuracy in disease prediction. Therefore, the outcomes from this paradigm are mostly cared in the experiment and can serve as an indicator of the quality of the generated dataset as well as the performance of the fine-tuned model.

- generation-only: This experimental design uses only the generated data for training, testing, and evaluation. Intended as a control setup, its design is aimed at quantifying the extent of overfitting attributable to data redundancy.

Training outcomes for both PLMs and LLMs are presented in Table 4. Upon comparing and analysing the prediction accuracy as presented in Table 4, the following observations are deduced:

Overall comparison *For each model, when trained using datasets of 100 cases, 200 cases, 500 cases, and 800 cases under each experimental setup, it becomes evident that the incorporation of generated data significantly enhances prediction accuracy. The trend indicates that as the data scales from 100 cases for each disease to 500 cases, there is a substantial improvement in predictive accuracy. Conversely, when this scale increases from 500 to 800 cases, the rate of improvement decelerates. To elucidate, during the LlaMa-7B training under the mixed-train-collect-test experiment, the accuracy surged by 0.63 when the dataset increased from 100 to 500 cases. However, in the transition from 500 to 800 cases, the accuracy increased by only 0.059. This enhancement suggests that data augmentation techniques exert a beneficial effect on downstream disease classification tasks, and the resulting augmented dataset exhibits a substantial quantity.*

In fields grappling with data sparsity, employing generation-based LLMs such as ChatGPT for data augmentation proves effective and merits exploration. Nonetheless, it is pivotal to acknowledge that the benefits of generated data exhibit diminishing returns. Beyond a certain threshold of generated data volume, the incremental improvements in downstream classification accuracy become relatively marginal. Furthermore, Table 4 consistently shows that BERT and BioBERT achieve higher classification accuracy than the LlaMa-7B model across all evaluated data volumes (100, 200, 500 and 800 cases). The primary explanation for this performance difference resides in the inherent mismatch between generative model architectures and discriminative tasks. While LLMs fundamentally operate by predicting the next token, classification necessitates precise differentiation between predefined categories. Capturing linguistic richness is central to generation, whereas classification demands exacting decision boundaries. Although the multiple-choice QA format constrained potential outputs, the fundamental discord between LLaMa's generative architecture and the discriminative requirements of the classification task likely imposed limitations on optimal performance. Consequently, the experimental results substantiate that traditional fine-tuning approaches, such as those employing BERT-based models, remain preferable for computationally sensitive scenarios requiring high-stakes decision-making with high precision demands.

Comparison of four experiment setups *To mitigate concerns related to overfitting and bias arising from the augmented dataset, this section establishes four experimental configurations encompassing various combinations of generated data and actual referral letters. The diminishing returns in data augmentation manifest primarily when available real referral letters are exhausted as generation prompts. This depletion of heterogeneous patterns reduces synthesized data diversity, yielding only marginal accuracy gains beyond the 100-case threshold. Table 4 presents the predictive accuracy of these setups across different models, revealing a consistent trend. Specifically, when the training dataset for each disease exceeds 100 cases, the predictive accuracy follows the order: mixed > generation − only > mixed − train − collect − test > gen − train − collect − test. This performance gradient is attributable to two factors: (1) the enhanced representational capacity derived from the retention of authentic clinical patterns in the "mixed" configuration, and (2) the susceptibility of validation sets containing synthetic samples to data leakage artifacts, which inflate reported accuracy. Consequently, subsequent analyses prioritize the "mixed-train-collect-test" and "gen-train-collect-test" setups to ensure evaluation integrity. This observed trend aligns with the design principles elucidated in Section 3.3.3.*

Table 4 Prediction accuracy of models on designed setups

Models	Experiment setups	100 cases	200 cases	500 cases	800 cases
LlaMa-7B	mixed	0.459	0.717	0.881	0.905
	mixed-train-collect-test	0.034	0.336	0.664	0.723
	generation-only	0.334	0.673	0.871	0.905
	gen-train-collect-test	0.200	0.228	0.524	0.592
BERT	mixed	0.902	0.963	0.974	0.981(F1:0.9805)
	mixed-train-collect-test	0.800	0.864	0.923	0.977(F1:0.9707)
	generation-only	0.918	0.957	0.977	-
	gen-train-collect-test	0.727	0.785	0.874	0.877
BioBERT	mixed	0.885	0.947	0.979	0.978
	mixed-train-collect-test	0.782	0.841	0.968	0.973
	gen-train-collect-test	0.710	0.802	0.894	0.932
LlaMa-13B	mixed			0.873	
	mixed-train-collect-test			0.621	
	generation-only			0.864	
	gen-train-collect-test			0.526	

Given that both "mixed" and "mixed-train-collect-test" setups incorporate actual referral letters, their performance warrants focused attention. Consequently, Fig. 4 illustrates the performance of these experimental setups on the BERT model, which demonstrated the highest classification accuracy among the four setups in the scenario involving 800 cases. Figure 4.a and (c) comprehensively visualize the learning dynamics encompassing training loss, validation accuracy, training accuracy and validation loss across epochs, while Fig. 4.b and (d) depict the corresponding confusion matrices quantifying diagnostic category performance.

Figure 4.a illustrates the performance of the mixed setup, depicting a noteworthy issue from epochs 0 to 20. During this period, the validation loss consistently surpasses the training loss, while the validation accuracy concurrently outperforms the training accuracy. This anomaly is indicative of a potential "data leakage" scenario, wherein certain generated data exhibit notable similarities and are allocated to both the training and validation datasets. Consequently, it is imperative to confine the inclusion of generated data solely to the training dataset for this disease classification task, mitigating the risk of data leakage". Conversely, the performance of the mixed-train-collect-test" setup, as depicted in Fig. 4.c, exhibits commendable performance, as evidenced by the coherent trends observed in loss and accuracy metrics for both training and validation datasets. This alignment is further corroborated by the confusion matrix presented in Fig. 4.d, which similarly indicated balanced category performance without detectable systematic bias. These collective observations suggest that the employed "mixed-train-collect-test" methodology adheres to a reasonable dispatch principle, yielding generated data of sufficiently high quality to significantly enhance disease classification accuracy as the generated training dataset size increases. Furthermore, the absence of systematic bias in both confusion matrices substantiates the effectiveness of the overall methodology involving data augmentation followed by fine-tuning of the encoder-based PLM. This robustness is quantitatively reinforced by the primary outcomes presented in Table 4, which include the F1 metric for this configuration. Consequently, the "mixed-train-collect-test" setup proves valuable for assessing the performance of various models in subsequent discussions related to disease classification.

Comparison of BERT and BioBERT *Excluding the "mixed" and "generation-only" configurations due to concerns of overfitting, analysis indicated that both BioBERT and BERT achieve similar prediction accuracy levels in the "gen-train-collect-test" and "mixed-train-collect-test" scenes, particularly when considering datasets with fewer than 100 cases. However, as the dataset size expands, BioBERT consistently surpasses BERT in performance. This superior performance can be attributed to the specialised training of BioBERT on extensive medical datasets, which imparts more domain-specific knowledge compared to BERT's broader, generalist training. Notably, within the "mixed-train-collect-test" configuration, BERT achieves a prediction accuracy of 0.977, marking the highest recorded result in this study.*

Comparison between BERT/BioBERT and LlaMa-7B *In evaluating the results of the "mixed-train-collect-test" and "gen-train-collect-test" predictions, this study observed that encoder-based PLMs such as BERT and BioBERT consistently outperformed decoder-based LLMs in free-text classification tasks. This superiority can be attributed to the fact that encoder-based models generate a probability distribution over a fixed category. Conversely, the decoder-based model's output, in terms of length, format, and other parameters, remains difficult to control. Even when disease categories are presented as options within the prompt text, managing model outputs remains a challenge. Furthermore, the hardware and training expenses associated with BERT/BioBERT are substantially less than those of LLMs. Therefore, for text classification tasks involving fixed categories, encoder-based PLMs appear to be a more efficient choice over LLMs.*

Comparison of LlaMa-7B with LlaMa-13B *Due to the dual A5000 graphic cards, it is unable to execute the LlaMa-13B model. Consequently, this study secured four A100 GPUs and fine-tuned the model using a dataset of 500 cases. The results indicate that the prediction accuracy of LlaMa-13B closely mirrors that of LlaMa-7B, with minor deviations of no more than 0.2 in any of the four experimental setups.*

Comparison of LlaMa-7B with ChatGPT and GPT-4 *Table 3 demonstrates that, among ChatGPT and GPT-4, the highest predictive accuracy is achieved through the few-shot learning approach of ChatGPT-4, resulting in an accuracy of 0.544. Disregarding the potential impact of overfitting, the results of LlaMa-7B in both the "mixed-train-collect-test" and "gen-train-collect-test" experiments were 0.723 and 0.592, respectively. These scores surpass those of GPT-4. Given the costs associated with the GPT-4 API and concerns related to patient data privacy, open-source LLM fine-tuning may present a more suitable alternative for downstream tasks reliant on patient data.*

4 Limitations

The dataset was Limited to 17 neurological conditions from a single UHD center, potentially introducing selection and spectrum biases due to insufficient representation of rare disorders or complex comorbidities. Consequently, pathway generalizability requires validation across comprehensive, multi-center datasets. Additionally, scaling fine-tuning to LLMs exceeding 7B parameters was precluded by hardware constraints, though encoder-based PLMs achieved viable performance without computational augmentation.

The feasibility of clinical deployment is currently under pilot evaluation using a BioBERT data-augmented model with selected neurologists; subsequent phases involving expanded datasets/models depend on initial results.

5 Conclusion

In this research, optimal strategies for assisting GPs through referral letter analysis were systematically investigated, with dual objectives of establishing clinical decision pathways and evaluating LLMs for disease classification. A novel in-context learning diagnostic framework was developed and validated using 439 anonymized referral letters, incorporating direct diagnosis and multiple-choice QA paradigms. Results demonstrated that the diagnostic accuracy of both ChatGPT and GPT-4 remained clinically insufficient for real-world implementation, necessitating alternative methodologies. To mitigate medical data scarcity constraints, structured data augmentation was implemented through ChatGPT-generated synthetic text, followed by rigorous fine-tuning of encoder-based PLMs (BERT/BioBERT) and decoder-based LLMs (LlaMa-7B/13B). Experimental validation confirmed that augmented data risks were effectively controlled via stratified sampling and regularization techniques, ensuring robust classification performance. The establishment of clinical translation pathways is substantiated by three pivotal findings: encoder-based PLMs consistently demonstrated superior performance over decoder-based LLMs in referral letter-based diagnostic tasks, while BERT delivered state-of-the-art classification accuracy (0.977) approaching clinical utility thresholds, and F1(0.9707) & confusion matrix analysis – confirmed absence of systematic diagnostic bias. This evidential robustness demonstrates consistently balanced performance across neurological categories, validating methodological reliability. These collective outcomes define a deployable clinical framework comprising three core components: LLM-mediated data augmentation for specialized diagnostic applications, prioritization of resource-efficient encoder-PLMs in real-time decision support systems, and implementation of a staged validation protocol beginning with neurologist-annotated BERT pilot studies before expanding to multicenter trials that broaden diagnostic scope to diverse conditions.

6 Supplementary information

Appendix A presents the examples of referral letters. Appendix B provides prompt examples for the direct diagnosis process. Appendix C illustrates prompt example of multiple-choice question and answering method.

Appendix A The examples of referral letters

Table 5 The examples of referral letters

Complaint text	Disease diagnosis
Dear Dr XXX, We would be grafeful if you could review Mr XXXX in one of your Rapid Acsess Neurology Clinic. He was seen on Ansty under Dr XXXX with a history of severe sudden onset headache followed by visual disturbances and unsteadyness within minutes. His visual disturbance is unusual, describing bilateral diplopia which persists as bi and monocular vision. Initially he also described very restriced tunnel vision that have steadily improved during his admission but still not returned to normal. On examination he has normal ocular movements, symmetrical tunnelling of the visual fields in both eyes. His pupillary reflexes and fundoscopy are normal. No other abnormal neurology was detected.	R56.9 - Unspecified convulsions
39 year old lady was involved in a road traffic accident about two weeks ago when her car was shunted from behind and she experienced some low back pain. She felt it was getting gradually better but in the last 48 hours the pain has got worse and has been associated with urinary incontinence and faecal incontinence. Getting some tingling in both legs.She had an MRI scan of her spine last night but it was completely clear. Her problem does not seem to be due to any spinal problem. Discussion with general surgeons: informed that her presentation would not constitute a general surgical emergency and that the patient should be referred to a pelvic floor surgeon via GP. Medical registrar advised that all back pain patoients are to reviewed by rheumatology team and no medical team imput is necessary. Discussed with rheumatology consutant: no need for rheumatology review, reffer to Neurology tweam.	M54.5 - Low back pain
17 year old with first episode of seizure PC - seizure at work 10-12-2017 morning - customer noted eyes roll back, became rigid, fell backwards, shaking arms and legs for 2 minutes. Tongue biting, no incontinence. 1 minute later, had another similar episode. Postictal phase - can't remember going to work. Later that night, attended party - had excess alcohol, friend stated knocked back whilst sitting on curb and hit her head on pavement, no immediate loss of consciousness.	G40.3 - Generalised idiopathic epilepsy and epileptic syndromes
We would be grateful if you could see this 35year old lady as a inpatient on Ansty. She has previously been seen by Dr XXXX with BIH and had a VP shunt inserted on the 9-3-18. She required adjustment to this on 23-3-18 as VP shunt was in subcutaneous fat instead on peritoneum. She has presented with a week history of worsening headache and neck pain with associated blurred vision. On examination she has bilateral papilloedema. CT and shunt series have been performed and reviewed by SOTON. Initially they suggested proceeding to LP which unfortunately was unsuccessful and Anaesthetic assistance has been requested on CEPOD. SOTON have been back in contact and suggest ophthalmology review first. We would be grateful for your assessment of this lady with her management and treatment.	G43.9 - Migraine, unspecified

Appendix B Prompt examples for the direct diagnosis process

Table 6 Prompt examples for the direct diagnosis process

Prompt type	Prompt examples
Zero-shot	Please provide only the ICD-10 code or the official name of the diagnosed disease as stated in the following referral letter. Referral letter: Presented with tonic clonic seizure, self-resolved lasted approximately 2-3 minutes Reports had no warning of episode, doesn't remember period after. Remembers coming in the ambulance. Now memory poor, increased confusion AMTS: 5-10. Not at baseline-daughter reports normally happens after seizure. Has had increased episodes of passing out Had similar episodes before, on sodium valproate Not been seen by neuro before BG: dementia ETOH excess, no withdrawal symptoms. Diagnosis:
One-shot	Please provide only the ICD-10 code or the official name of the diagnosed disease as stated in the following referral letter. Referral letter: Presented with tonic clonic seizure, self-resolved lasted approximately 2-3 minutes Reports had no warning of episode, doesn't remember period after. Remembers coming in the ambulance. Now memory poor, increased confusion AMTS: 5-10. Not at baseline-daughter reports normally happens after seizure. Has had increased episodes of passing out Had similar episodes before, on sodium valproate Not been seen by neuro before BG: dementia ETOH excess, no withdrawal symptoms. Diagnosis: G409 - Epilepsy, unspecified Referral letter: Presented with tonic clonic seizure, self-resolved lasted approximately 2-3 minutes Reports had no warning of episode, doesn't remember period after. Remembers coming in the ambulance. Now memory poor, increased confusion AMTS: 5-10. Not at baseline-daughter reports normally happens after seizure. Has had increased episodes of passing out Had similar episodes before, on sodium valproate Not been seen by neuro before BG: dementia ETOH excess, no withdrawal symptoms. Diagnosis:
One-shot-symp	Please provide only the ICD-10 code or the official name of the diagnosed disease as stated in the following referral letter. Referral letter: Presented with tonic clonic seizure, self-resolved lasted approximately 2-3 minutes Reports had no warning of episode, doesn't remember period after. Remembers coming in the ambulance. Now memory poor, increased confusion AMTS: 5-10. Not at baseline-daughter reports normally happens after seizure. Has had increased episodes of passing out Had similar episodes before, on sodium valproate Not been seen by neuro before BG: dementia ETOH excess, no withdrawal symptoms. [SEP]tonic clonic seizure [SEP] memory poor [SEP] confusion [SEP] seizure [SEP] passing out [SEP] dementia. Diagnosis: G409 - Epilepsy, unspecified. Referral letter: Chest pain and collapse x2 (once in ED, hit head on 1st seizure) Impression - Generalised seizures secondary to old stroke. Please can you come and assess the patient and advise on which Antiepileptic medication can be started [SEP]Chest pain [SEP] collapse [SEP] hit head [SEP] seizure [SEP] Generalised seizures [SEP] stroke. Diagnosis:

Appendix C Prompt example of a multiple-choice question and answering method

Table 7 Prompt example of a multiple-choice question and answering method

Prompt type	Prompt examples
Zero-shot	Definition: For this task, we request that you make a diagnosis by examining the provided Referral letter.
	Referral letter: Dear Dr XXX, We would be grateful if you could review Mr XXXX in one of your Rapid Access Neurology Clinic. He was seen on Ansty under Dr XXX with a history of severe sudden onset headache followed by visual disturbances and unsteadiness within minutes. His visual disturbance is unusual, describing bilateral diplopia which persists as bi and monocular vision. Initially he also described very restricted tunnel vision that have steadily improved during his admission but still not returned to normal. On examination he has normal ocular movements, symmetrical tunneling of the visual fields in both eyes. His pupillary reflexes and fundoscopy are normal. No other abnormal neurology was detected.
	Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalised idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine, unspecified (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 - Pneumonia, unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 - Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions.
	Output:
One-shot	Definition: For this task, we request that you make a diagnosis by examining the provided Referral letter.
	Referral letter: This gentleman was admitted with SUO after a recent discharge from RBH for rehab 2-52. She has a background of RA and polyarthralgia, which has been worsening over 6-12. Had transient R sided weakness sept 2018. We would appreciate a neurology opinion on her globally increased tone and decreased mobility. Thank you.
	Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalised idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine, unspecified (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 - Pneumonia, unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 - Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions.
	Output: A
	Referral letter: Mr XXXX, 72 year-old gentleman, was admitted post-radical chemo-radiotherapy for SCC oesophagus with reduced mobility and poor oral intake. On assessment we are concerned that some of these symptoms may be related to possible Parkinsonism.
	Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalised idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine, unspecified
	(F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 - Pneumonia, unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 - Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collaps (Q) R56.9 - Unspecified convulsions.
Output:	
Few-shot	Definition: For this task, we request that you make a diagnosis by examining the provided Referral letter.
	Referral letter: This gentleman was admitted with SUO after a recent discharge from RBH for rehab 2-52. She has a background of RA and polyarthralgia, which has been worsening over 6-12. Had transient R sided weakness sept 2018. We would appreciate a neurology opinion on her globally increased tone and decreased mobility. Thank you.
	Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collaps (Q) R56.9 - Unspecified convulsions.
	Output: A
	Referral letter: Repat from SGH: Previous subarachnoid hemorrhage 16th September 2014. Aneurysm was coiled in SGH on the 17th September. She developed a focal seizure and decreased GCS on the 22nd September and there was no sign of ischaemia or infarct on the repeat CT. She was treated with nimodipine and phenytoin. Phenytoin has now stopped. A repeat CT scan on the 2nd October showed a small region of right frontal low density likely due to vasospasm. Mrs. XXXX also developed hospital acquired pneumonia and improved on chloramphenicol.

Table 7 (continued)

Prompt type	Prompt examples
	Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions. Output: G Referral letter: Admitted acutely as thrombolysis call with right sided weakness in her arm and leg and speech disturbance. Thrombolysis not performed as diagnosis unclear. CTH and MRI brain normal. Symptoms are ongoing and are thought to be functional. We wonder if there are any further investigations we should be doing to aid a diagnosis or if she needs neurological follow up. Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions. Output: M Referral letter: Dear Dr XXXX, We would be grateful if you could review this 50-year-old Gentleman who you have met previously. He was admitted on the 12-4-18 with severe headache post head injury 2-52 prior. He has had a normal MRA and LP but is struggling with his headache despite regular analgesia, fluids and amitriptyline. Many thanks Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions. Output: O Referral letter: Admitted after fall with loss of consciousness, thought likely secondary to postural hypotension. Reports 2 falls per day. Inflammatory markers raised on admission but settled after treatment for aspiration pneumonia. Normally lives fairly independently at home, no package of care. PMHx AF, CVA 2017 (two tiny sites of hemorrhage involving the right cerebellar hemisphere and corpus callosum), thalassemia trait, portal hypertension Poor oral intake prior to admission, with associated weight loss on background of low BMI previously. BMI 17. Reviewed by speech and language therapy who know her from previously. Found to have severe oropharyngeal dysphagia and advised for NBM and NG feed. Felt to be significantly worse than during previous admissions; cause of deterioration unclear. Fatigability noted over the course of assessment. CT head and MRI for further CVE demonstrated no evidence of infarct or hemorrhage. Small vessel disease including the pons was noted. Speech is of a nasal quality, with ataxic dysarthria. Awaiting video fluoroscopy Friday, and AChR antibodies. We would be grateful for your opinion as to whether her symptoms could all be explained by previous stroke with small vessel disease of the brainstem, or whether this might represent an alternative etiology such as myasthenia or bulbar palsy. Many thanks. Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions. Output: P Referral letter: Mr XXXX, 72 year-old gentleman, was admitted post-radical chemo-radiotherapy for SCC esophagus with reduced mobility and poor oral intake. On assessment we are concerned that some of these symptoms may be related to possible Parkinsonism. Question: Which of the following does this referral letter refer to? Please only output the option. (A) A41.9 - Sepsis, unspecified organism (B) G35 - Multiple sclerosis (C) G40.3 - Generalized idiopathic epilepsy and epileptic syndromes (D) G40.9 - Epilepsy, unspecified (E) G43.9 - Migraine (F) G61.0 - Guillain-Barre syndrome (G) I60.9 - Nontraumatic subarachnoid hemorrhage, unspecified (H) I63.9 - Cerebral infarction, unspecified (I) J18.9 – Pneumonia unspecified organism (J) J69.0 - Pneumonitis due to inhalation of food and vomit (K) M54.5 - Low back pain (L) N39.0 - Urinary tract infection, site not specified (M) R29.8 – Other symptoms and signs involving the nervous and musculoskeletal systems (N) R41.0 - Disorientation, unspecified (O) R51 - Headache (P) R55 - Syncope and collapse (Q) R56.9 - Unspecified convulsions. Output:

Acknowledgements During the research process and paper preparation, I received selfless assistance from my supervisor and numerous colleagues. I am immensely grateful for their invaluable contributions.

Author Contributions Project administration, supervision and conceptualization: RP, XY, XK, HL and JZ; Methodology, experiment, visualization and writing-original draft preparation: RW, AR and TL; Writing–review and editing: XY, RW, AR, HL and XK; Funding acquisition HL; Validation, formal analysis and data curation: RW, XY and RP. All authors read and approved the final manuscript.

Funding This research has been supported by the Neuravatar Project funded by High Education Innovation Fund (UK-HEIF), match funded PhD scholarship from Bournemouth University and Shandong Hengdao Ruyi Digital Communication Ltd, the Cloud based Virtual Production Project funded by the Arts and Humanities Research Council (UK-AHRC Ref: AH/W009323/1).

Data Availability The generated referral letters can be found in the github link: https://github.com/ruibin-wang/solution-assisting-GPs.git.

Code Availability The collected dataset utilized in the study is currently inaccessible as the collaborating doctor has no plan to share the data, citing complicated policies that must be followed for releasing a dataset. The sample of referral letters can be found at Appendix A-Appendix C. Besides, the code and generated referral letters can be found in the github link: https://github.com/ruibin-wang/solution-assisting-GPs.git.

Materials Availability Not applicable.

Declarations

Competing interests The authors declare no competing interests.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Lewenberg Y, Bachrach Y, Paquet U, Rosenschein J (2017) Knowing what to ask: A bayesian active learning approach to the surveying problem. In: Proceedings of the AAAI conference on artificial intelligence, vol 31
2. Shim H, Hwang SJ, Yang E (2018) Joint active feature acquisition and classification with variable-size set encoding. Adv Neural Inf Process Syst 31
3. Wei Z, Liu Q, Peng B, Tou H, Chen T, Huang XJ, Wong KF, Dai, X.: Task-oriented dialogue system for automatic diagnosis. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) pp. 201–207
4. Xu L, Zhou Q, Gong K, Liang X, Tang J, Lin L (2018) End-to-end knowledge-routed relational dialogue system for automatic diagnosis. Proceedings of the AAAI conference on artificial intelligence 33:7346–7353
5. Wei J, Zou K (2019) Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196
6. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
7. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2019) Bart: Denoising sequence-to-sequence pre-training for natural language generation translation and comprehension. arXiv preprint arXiv:1910.13461
8. Wu X, Lv S, Zang L, Han J, Hu S (2019) Conditional bert contextual augmentation. In: Computational science–ICCS 2019: 19th international conference faro Portugal June 12–14 2019 Proceedings Part IV 19, Springer, pp 84–95
9. OpenAI (2023) Gpt-4 technical report. ArXiv. abs/2303.08774
10. Bayer M, Kaufhold MA, Reuter C (2022) A survey on data augmentation for text classification. ACM Comput Surv 55(7):1–39
11. Feng SY, Gangal V, Wei J, Chandar S, Vosoughi S, Mitamura T, Hovy, E (2021) A survey of data augmentation approaches for nlp. arXiv preprint arXiv:2105.03075
12. Taori R, Gulrajani I, Zhang T, Dubois Y, Li X, Guestrin C, Liang, P, Hashimoto TB (2023) Alpaca: A strong replicable instruction-following model. Stanford Center Res Found Models 3(6)7. https://crfm.stanford.edu/2023/03/13/alpaca. html
13. Dai H, Liu Z, Liao W, Huang X, Wu Z, Zhao L, Liu W, Liu N, Li, S, Zhu D et al (2023) Chataug: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:2302.13007
14. Nori H, King N, McKinney SM, Carignan D, Horvitz E (2023) Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375
15. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, Scales, N, Tanwani A, Cole-Lewis H, Pfohl S et al (2022) Large language models encode clinical knowledge. arXiv preprint arXiv:2212.13138
16. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
17. Doğan RI, Leaman R, Lu Z (2014) Ncbi disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform 47:1–10
18. Li J, Sun Y, Johnson RJ, Sciaky D, Wei CH, Leaman R, Davis, AP, Mattingly CJ, Wiegers TC, Lu Z (2016) Biocreative v cdr task corpus: A resource for chemical disease relation extraction. Database 2016
19. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma'ayan A (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database
20. Van Mulligen EM, Fourrier-Reglat A, Gurwitz D, Molokhia M, Nieto A, Trifiro G, Kors JA, Furlong LI (2012) The eu-adr corpus: Annotated drugs, diseases targets and their relationships. J Biomed Inform 45(5):879–884
21. Li T, Shetty S, Kamath A, Jaiswal A, Jiang X, Ding Y, Kim Y (2023) Cancergpt: Few-shot drug pair synergy prediction using large pre-trained language models. arXiv preprint arXiv:2304.10946
22. Edwards C.N, Naik A, Khot T, Burke M.D, Ji H, Hope T.: Synergpt: In-context learning for personalized drug synergy prediction and drug design. bioRxiv 2023–07 (2023)
23. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix, T, Rozière B, Goyal N, Hambro E, Azhar F, et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. (2023)
24. Zeng A, Liu X, Du Z, Wang Z, Lai H, Ding M, Yang Z, Xu Y, Zheng W, Xia X et al (2022) Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414

25. Wang R, Jayathunge K, Page R, Li H, Zhang JJ, Yang X (2024) Hybrid architecture based intelligent diagnosis assistant for gp. BMC Med Inform Decis Mak 24(1):15
26. Ubani S, Polat SO, Nielsen R (2023) Zeroshotdataaug: Generating and augmenting training data with chatgpt. arXiv preprint arXiv:2304.14334
27. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240
28. Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. MIT Press. http://www.deeplearningbook.org
29. Sajjadi MSM, Oreshkin B (2018) Accuracy-recall tradeoffs in generative models. In: International conference on machine learning (ICML), pp. 811–820
30. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2021) Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685
31. Mangrulkar S, Gugger S, Debut L, Belkada Y, Paul S, Bossan B (2022) PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft