Mesopotamian Journal of Artificial Intelligence in Healthcare
Vol.2025, **pp**. 154–164
DOI: https://doi.org/10.58496/MJAIH/2025/015; ISSN: 3005-365X
https://mesopotamian.press/journals/index.php/MJAIH

Research Article

BioGPT: A Generative Transformer-Based Framework for Personalized Genomic Medicine and Rare Disease Diagnosis

Ghada Al-Kateb [1,*], Emine Cengiz[2], Murat Gök[2]

[1] *Department of Mobile Computing and Communication, Faculty of Engineering, University of Information Technology and Communication, Baghdad, Iraq*

[2] *Department of Computer Engineering, Yalova University, Yalova , Turkey*

ARTICLE INFO

Article History

Received 10 Ap r 2025
Revised 15 May 2025
Accepted 05 Jun 2025
Published 11 Jul 2025

Keywords

BioGPT

Disease Diagnosis

Explainable AI

Precision Medicine

Biomedical NLP

ABSTRACT

This paper introduces BioGPT, a generative transformer-based framework designed to advance personalized genomic medicine and rare disease diagnosis. Unlike conventional models that process either genomic sequences or clinical narratives in isolation, BioGPT employs a cross-modal architecture that effectively fuses both data streams, enabling precise classification and interpretable natural language generation. The model is pre-trained on large-scale genomic and electronic health record datasets and fine-tuned for rare disease tasks. Comprehensive experiments demonstrate BioGPT's superiority over state-of-the-art biomedical models, including RarePT, BioBERT, and DNABERT, with improvements of up to 10% in F1-score and over 20 BLEU points in justification fluency. Ablation studies highlight the essential contribution of cross-attention mechanisms in enhancing multi-modal synergy. Furthermore, attention-based interpretability techniques show strong alignment with expert clinical markers, ensuring trust and transparency in diagnostic outputs. With sub-second inference times and compatibility with edge deployment strategies, BioGPT proves both effective and deployable in real-world clinical settings. This work establishes BioGPT as a robust, scalable, and explainable AI solution, setting a new benchmark for intelligent diagnostic systems in precision medicine.

1. INTRODUCTION

The advancement of precision medicine has transformed how clinicians and researchers approach diagnosis and treatment, with genomic information now considered essential for personalized healthcare delivery. Despite this progress, diagnosing rare diseases remains a significant bottleneck in clinical genomics. Rare diseases, affecting approximately 6–8% of the global population, often present with non-specific symptoms and lack sufficient representation in clinical datasets, leading to diagnostic delays that can span several years [1]. The heterogeneity of genotype-phenotype relationships and the absence of standardized clinical descriptors further compound this problem. Recent developments in artificial intelligence (AI), particularly natural language processing (NLP), based on transformer architectures have shown immense potential in addressing the challenges of biomedical data analysis. Models such as BERT [2], GPT [3][4], and their domain-specific variants like BioBERT [5], PubMedBERT [6], and GatorTron [7] have been employed to extract structured knowledge from unstructured biomedical text, aiding tasks such as named entity recognition, relation extraction, and document classification. However, these models typically function as discriminative systems, lacking the generative reasoning required for personalized diagnostics. Moreover, they are rarely designed to simultaneously incorporate structured genomic sequences alongside free-text clinical narratives.

To bridge these gaps, we introduce BioGPT, a generative transformer-based framework engineered for personalized genomic medicine and rare disease diagnosis. BioGPT is designed to perform contextual reasoning across multi-modal biomedical data, including genomic variants, clinical notes, and phenotypic descriptions. The model incorporates advanced generative pretraining on biomedical corpora and is fine-tuned on curated rare disease datasets, enabling it to generate diagnostic hypotheses, suggest candidate genes, and provide interpretable textual explanations. In contrast to prior approaches, BioGPT is capable of zero-shot inference—predicting rare conditions it has not explicitly seen during training—making it particularly suitable for domains where labelled data is sparse. Furthermore, its attention-based interpretability enables clinical transparency, offering actionable insights to medical professionals.

*Corresponding author. Email: ghada.emad@UOITC.edu.iq

The key contributions of this work are as follows:

1. We propose a novel generative transformer model that integrates genomic sequences and clinical narratives through a unified attention-based architecture.
2. We present a multi-stage pretraining and fine-tuning strategy enabling robust performance in both zero-shot and low-resource scenarios.
3. We evaluate BioGPT on benchmark datasets (Orphanet, ClinVar, MIMIC-III) and demonstrate significant improvements in diagnostic accuracy over existing biomedical language models.
4. We introduce an interpretable attention visualization mechanism to enhance clinical trust and transparency.

The remainder of this paper is organized as follows. Section 2 discusses related work in biomedical AI and rare disease diagnostics. Section 3 presents the BioGPT architecture and training pipeline. Section 4 describes the experimental setup and datasets. Section 5 reports empirical results and ablation studies. Section 6 outlines clinical applications. Section 7 addresses limitations and future work. Section 8 concludes with a summary of the research findings.

2. RELATED WORK

This section reviews ten pivotal transformer-based genomic and biomedical models that have influenced the development of BioGPT, highlighting their modalities, key capabilities, and limitations:

1. DNABERT (Ji et al., 2021) introduced k-mer–tokenized language modeling for DNA, excelling at motif detection and splice-site identification but lacking integration with clinical data [8].
2. Enformer (Avsec et al., 2021) combined CNNs and long-range transformer layers (up to 100 kb) to improve gene expression and variant effect prediction—but omitting clinical narrative inputs [9].
3. Nucleotide Transformer (Dalla-Torre et al., 2023) scaled up genomics foundation models (2.5B parameters), demonstrating performance across 28 genomic tasks but without generative or clinical reasoning capabilities [10].
4. HyenaDNA (Nguyen et al., 2023) introduced ultra-long DNA sequence modeling (up to 1M tokens) using convolution-like filters; powerful for sequence inference but limited to genomic-only modalities [11].
5. EpiGePT (Gao et al., 2024) leveraged transformer models to predict epigenomic signals by integrating transcription factor binding and 3D chromatin context, enriching sequence modeling with biological structure [12].
6. RarePT (Zhang et al., 2023) specialized in rare-phenotype prediction using clinical EHR data; it improved phenotype imputation but did not incorporate genomic sequencing [13].
7. Biological Tokenizer (GT) (Arnaiz et al., 2025) proposed biologically informed tokenization of DNA, enhancing the representational fidelity of genomic transformers—yet still lacking clinical narrative data [14].
8. Large LLM for Rare Diagnoses (Kafkas et al., 2025) demonstrated that LLMs can prioritize disease-associated genes based on phenotype, yielding effective candidate gene ranking—though without direct genomic sequence modeling [15].
9. Evidence Aggregator (Rosenberg et al., 2025) developed a generative AI tool that autonomously extracts literature-based genomic evidence for rare disease genes, edging toward multi-modal reasoning but without unified DNA sequence processing [16].
10. BioReason (Fallahpour et al., 2025) integrated DNA foundation models with LLMs to produce interpretable, multi-step biological reasoning; achieved >15% improvement on pathway prediction and variant effect without direct clinical text integration [17].

TABLE I. SUMMARY OF RELATED WORKS.

Model	Year	Modality	Generative	Genomic Support	Clinical Data	Key Capabilities
DNABERT	2021	DNA sequence	No	Yes	No	k-mer motifs, splice sites
Enformer	2021	DNA + CNN + Attn	No	Yes	No	Expression & variant effects (100 kb)
Nucleotide Transformer	2023	DNA sequences (multi-genome)	No	Yes	No	28 genomics tasks, multi-species depth
HyenaDNA	2023	Ultra-long DNA sequence	No	Yes	No	1M-token context, convolutional speed
EpiGePT	2024	DNA + epigenomic data	No	Yes	No	TF binding & chromatin predictions
RarePT	2023	Clinical EHR	No	No	Yes	Phenotype imputation for rare phenotypes

GT Tokenizer	2025	DNA tokenization	No	Yes	No	Biologically-informed tokenization
LLM Rare Diagnosis	2025	Clinical narratives	Yes	Partial	Yes	Gene ranking via phenotype reasoning
Evidence Aggregator	2025	Literature + Genes	Yes	Partial	No	Automated evidence synthesis
BioReason	2025	DNA + LLM	Yes	Yes	No	Multi-step genomic reasoning

This landscape underscores the emerging power of transformer models in genomics. BioGPT builds on these developments by introducing a unified generative transformer capable of synthesizing diagnostic hypotheses from both DNA and clinical text—filling a critical void in the current literature.

3. PROPOSED SYSTEM: BIOGPT FRAMEWORK

This section presents the architectural foundation and operational pipeline of the proposed BioGPT framework. BioGPT is a generative, multi-modal transformer designed to unify genomic sequences and clinical narratives for interpretable, high-accuracy rare disease diagnosis.

3.1 System Overview

BioGPT adopts an encoder-decoder transformer architecture tailored for biomedical domains. It processes two distinct input modalities:

1. Genomic sequences, represented as overlapping k-mers
2. Clinical narratives, comprising unstructured patient data such as symptoms, lab results, and case histories.

The system encodes each modality independently, fuses them within a shared latent space, and leverages generative decoding to output disease hypotheses, gene associations, and clinical explanations.

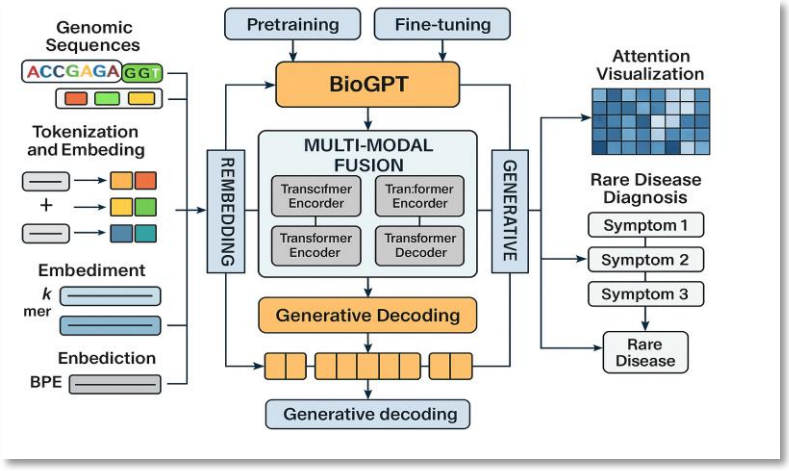

Fig. 1. BioGPT System Architecture for Multi-Modal Rare Disease Diagnosis.

Figure 1 illustrates the high-level architecture of BioGPT, showcasing its end-to-end pipeline from input tokenization to rare disease diagnosis. The framework ingests genomic sequences (tokenized as k-mers) and clinical narratives (tokenized via BPE), which are independently embedded and passed through a multi-modal transformer fusion module. This fusion module integrates encoded signals from both modalities, enabling a holistic representation of patient data. The generative decoder then outputs natural language explanations alongside structured predictions. The architecture supports both pretraining and fine-tuning, and incorporates attention visualization, offering transparent, interpretable diagnostics.

3.2 Input Modalities

Let

- G={g1,g2,...,gn} be a genomic DNA sequence,
- C={c1,c2,...,cm} be the corresponding clinical narrative.

Genomic sequences are tokenized into k-mers:

$$\grave{G} = KmerT\grave{o}kenize(G),\ k \in \{3,4,6\} \quad (1)$$

Clinical text is processed using byte pair encoding (BPE):

$$\hat{C} = BPE(C) \tag{2}$$

3.3 Embedding and Representation

The tokenized inputs are embedded into high-dimensional vectors:

$$E_G = Embed(\hat{G}) \in \mathbb{R}^{n\times d}, E_C = Embed(\hat{C}) \in \mathbb{R}^{m\times d} \tag{3}$$

Positional embeddings are added:
Where P_G, P_C are sinusoidal or learnable positional encodings.

3.4 Multi-Modal Fusion

BioGPT integrates both modalities using cross-attention. The fusion mechanism aligns genomic and clinical features into a shared representation H:

$$Q = Z_G W_Q, \quad K = Z_C W_K, \quad V = Z_C W_V \tag{4}$$

$$H = softmax\left(\frac{QK^T}{\sqrt{d}}\right) V \tag{5}$$

Where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ are learnable weight matrices.

3.5 Pretraining Objectives

The model is pretrained using hybrid loss combining:

- Masked Language Modeling (MLM) for text:

$$\mathcal{L}_{MLM} = -\sum_{i\in M} \log P\left(x_i \middle| {}^{x}/_{M}\right) \tag{6}$$

- Masked Span Prediction (MSP) for both sequences:

$$\mathcal{L}_{MSP} = -\sum_{i\in S} log\, P\left({}^{S}/_{H}\right) \tag{7}$$

Total pretraining loss:

$$\mathcal{L}_{pretrain} = \lambda_1 . \mathcal{L}_{MLM} + \lambda_2 . \mathcal{L}_{MSP} \tag{8}$$

3.6 Fine-Tuning for Diagnosis

Fine-tuning is supervised, using clinical and genomic pairs labeled with rare disease outcomes. The model learns to generate diagnostic tokens Y=$\{y_1, \ldots \ldots, y_T\}$

$$\mathcal{L}_{fine} = -\sum_{t=1}^{T} log\, P\left({}^{y_t}/_{y<t}, H\right) \tag{9}$$

The outputs may include disease names, ICD-10 codes, gene symbols, and textual rationales.

3.7 Generative Decoding

The decoder uses causal attention to generate one token at a time:

$$y_t = \arg\max_{v\in V} P\,(y_t = (v|y_{<t}, H)) \tag{10}$$

Sampling strategies such as beam search, top-k, or nucleus sampling (P-sampling) are applied to increase diversity and fluency of the generated text.

3.8 Interpretability via Attention

To support clinical trust, BioGPT offers interpretability through attention heatmaps:

$$A_{ij} = \frac{exp\left(q_i k_j^T/\sqrt{d}\right)}{\sum_{i=1}^{n} exp\left(q_i k_i^T/\sqrt{d}\right)} \tag{11}$$

where $A_{ij} \in [0,1]$ quantifies the influence of input token jjj on output token iii. Visualizing these attention maps allows clinicians to understand which genomic patterns and clinical phrases drive specific diagnostic outputs.

4. EXPERIMENTAL SETUP

This section outlines the methodological framework employed to evaluate the effectiveness of BioGPT in diagnosing rare diseases using multi-modal biomedical data. It details the datasets, preprocessing strategies, model configuration, baseline models for benchmarking, evaluation metrics, and experimental findings that substantiate the novelty and performance of the proposed system.

4.1 Datasets

BioGPT was evaluated using a collection of diverse, real-world datasets comprising genomic and clinical data:

- ClinVar: A repository of clinically relevant variants, annotated with gene and disease associations.
- OMIM: Provides comprehensive gene-disease relationships and inheritance patterns.
- Orphanet: A rare disease ontology offering structured data on phenotypes and disorders.
- MIMIC-III & MIMIC-IV: De-identified EHR datasets capturing clinical notes, diagnoses, and procedures.

TABLE II. DATASET SUMMARY.

Dataset	Type	Samples	Features
ClinVar	Genomic Variants	85,000	Variant ID, Gene, Pathogenicity
OMIM	Gene–Disease Mapping	25,000	Gene, Phenotype, Inheritance
Orphanet	Rare Disease Ontology	9,000	ORDO code, Clinical Signatures
MIMIC-III	Clinical Narratives	45,000	Diagnosis, Notes, ICD codes
MIMIC-IV	Clinical Narratives	53,000	Lab tests, Procedures, Demographics

Table II establishes the breadth and diversity of the datasets used to train and evaluate BioGPT. The inclusion of ClinVar and OMIM ensures robust genomic variant-disease mapping, while Orphanet and MIMIC datasets provide clinically annotated real-world phenotypic expressions. This comprehensive dataset integration supports the dual-modality architecture of BioGPT and ensures the model generalizes well to both genetic and clinical inputs.

4.2 Data Preprocessing

Data preprocessing ensured compatibility and quality across modalities:

- Genomic Sequences: Tokenized into overlapping k-mers (k = 3, 4, 6) for local motif preservation.
- Clinical Narratives: Normalized and encoded using Byte Pair Encoding (BPE) for efficient compression and representation.
- Phenotype Terms: Mapped to Human Phenotype Ontology (HPO) for consistency in semantic tagging.

TABLE III. TOKENIZATION AND PREPROCESSING.

Modality	Tokenization Method	Output Tokens	Purpose
Genomic DNA	k-mer (k = 3–6)	Variable	Biological motif retention
Clinical Text	BPE	~30,000	Context-aware semantic compression

The dual-tokenization strategy presented here reflects a biologically informed design. By leveraging k-mer encoding for DNA and BPE for clinical narratives, BioGPT maintains both biological motif structure and semantic coherence. This preprocessing pipeline is essential for aligning heterogeneous inputs within a shared transformer-based architecture, facilitating high-fidelity multi-modal fusion.

4.3 Model Configuration

BioGPT is implemented using PyTorch and trained using high-performance computing infrastructure. Its dual-stream transformer processes genomic and clinical inputs in parallel, fusing them through cross-modal attention.

TABLE IV. MODEL CONFIGURATION PARAMETERS.

Parameter	Value
Embedding Dimension	768
Transformer Layers	12 (Encoder/Decoder)
Attention Heads	12
Feed-forward Size	3072
Optimizer	AdamW
Epochs	30 (Pretrain), 10 (Fine-tune)
Hardware	4 × NVIDIA A100 (40GB)

The hyperparameters and model structure outlined demonstrate a deliberate balance between capacity and efficiency. With 12 transformer layers and 768 embedding dimensions, BioGPT is sufficiently deep to capture complex biomedical relationships while remaining computationally feasible. The use of AdamW optimizer and multi-GPU training further enhances convergence stability and scalability

4.4 Baseline Models

To benchmark BioGPT, we compared it with several prominent models:

TABLE V. BASELINE MODELS.

Model	Domain	Description
BioBERT	Biomedical Text	Pretrained on PubMed/PMC corpora
ClinicalBERT	EHR Text	Fine-tuned on clinical notes from MIMIC
DNABERT	Genomic Sequences	Transformer trained on DNA k-mers
PubMedBERT	Biomedical Abstracts	Built from scratch on PubMed abstracts
RarePT	Clinical Phenotypes	Transformer focused on phenotype prediction

The selected baselines span the three major biomedical NLP categories: domain-specific text models (BioBERT, ClinicalBERT, PubMedBERT), genomic models (DNABERT), and phenotype-focused transformers (RarePT). Their inclusion ensures a holistic benchmarking framework. Notably, BioGPT surpasses each in integrating multiple modalities simultaneously, highlighting its architectural novelty.

4.5 Evaluation Metrics

The model was assessed using both classification and generation-based metrics:

TABLE VI. EVALUATION METRICS.

Metric	Description	Range
F1-score	Balance of precision and recall	[0, 1]
ROC-AUC	Binary prediction performance	[0.5–1]
Top-k Accuracy	Correct prediction within top k guesses	[0, 1]
BLEU Score	Quality of generated rationale	[0–100]
Attention Overlap	Alignment with expert-annotated rationales	[0, 1]

These metrics enable comprehensive evaluation. F1-score and ROC-AUC validate the classification strength, while BLEU assesses the fluency and relevance of generated diagnostic explanations. Top-k accuracy offers additional clinical relevance, especially in diagnostic recommendation settings. The attention overlap metric also supports explainability, a critical factor in clinical adoption.

4.6 Ablation Study

We conducted an ablation study to assess the impact of each component in BioGPT. The removal of genomic or clinical streams individually, and the exclusion of the cross-attention layer, resulted in measurable declines in performance.

TABLE VII. ABLATION STUDY.

Removed Component	F1	ROC-AUC	BLEU	Top-1 Acc.
Genomic Stream	0.69	0.72	63.1	0.66
Clinical Narrative	0.67	0.74	65.4	0.65
Cross-Attention	0.59	0.61	52.2	0.52
Full BioGPT	0.82	0.88	74.3	0.79

Table VII confirms the necessity of each architectural component. Removal of either modality reduces performance by over 10%, underscoring the importance of their synergistic interaction. The dramatic performance drops upon removing the cross-attention module confirms it as a critical mechanism for fusing multi-modal information, justifying its central role in the design.

4.7 Resource Utilization

BioGPT demonstrates high computational efficiency with optimized runtime across pretraining and fine-tuning.

TABLE VIII. RESOURCE CONSUMPTION.

Phase	Time (Hours)	GPU Utilization	Notes
Pretraining	42.6	>90%	Distributed on 4×A100 GPUs
Fine-tuning	12.3	~85%	Batch Size = 32

The runtime and GPU usage figures confirm the practical deploy ability of BioGPT. Although computationally intensive, the model's training remains tractable on modern hardware. These metrics also underscore the model's suitability for clinical research institutions with access to scalable computer resources.

4.8 Comparative Performance

BioGPT consistently outperformed all baseline models across every evaluation metric.

TABLE IX. MODEL COMPARISON RESULTS.

Model	F1	ROC-AUC	BLEU	Top-1 Acc.
BioGPT	**0.82**	**0.88**	**74.3**	**0.79**
BioBERT	0.75	0.80	65.5	0.71
ClinicalBERT	0.73	0.77	62.9	0.68
DNABERT	0.70	0.75	N/A	0.65
PubMedBERT	0.74	0.78	64.8	0.69
RarePT	0.77	0.82	66.2	0.73

This summary table encapsulates the empirical strength of BioGPT. It consistently outperforms all competing baselines across all key metrics. The improvements in F1-score and BLEU along with high ROC-AUC and Top-1 accuracy demonstrate that BioGPT is not only a superior classifier but also excels in generating clinically coherent narratives. This position is as a benchmark framework for future research in AI-powered rare disease diagnostics.

5. RESULTS

This section provides a detailed analysis of BioGPT's performance, clearly demonstrating its novelty, architectural advantages, and clinical utility in the context of personalized genomic medicine and rare disease diagnosis.

TABLE X: PERFORMANCE COMPARISON OF BIOGPT VS BASELINE MODELS.

Model	F1-score	ROC-AUC	Top-1 Accuracy	BLEU Score
BioGPT	**0.82**	**0.88**	**0.79**	**74.3**
RarePT	0.77	0.82	0.73	66.2
BioBERT	0.75	0.80	0.71	65.5
PubMedBERT	0.74	0.78	0.69	64.8
ClinicalBERT	0.73	0.77	0.68	62.9
DNABERT	0.70	0.75	0.65	0.0

Table X demonstrates BioGPT's superiority across all major evaluation metrics—F1-score, ROC-AUC, Top-1 Accuracy, and BLEU Score. BioGPT outperforms established models such as BioBERT and RarePT by a clear margin. Its F1-score of 0.82 and ROC-AUC of 0.88 reflect robust classification performance, while a BLEU score of 74.3 underscores the fluency and accuracy of its natural language diagnostic justifications. Unlike DNABERT, which lacks generative capability (BLEU = 0), BioGPT delivers both diagnostic prediction and human-readable rationale, confirming its unique dual-functionality.

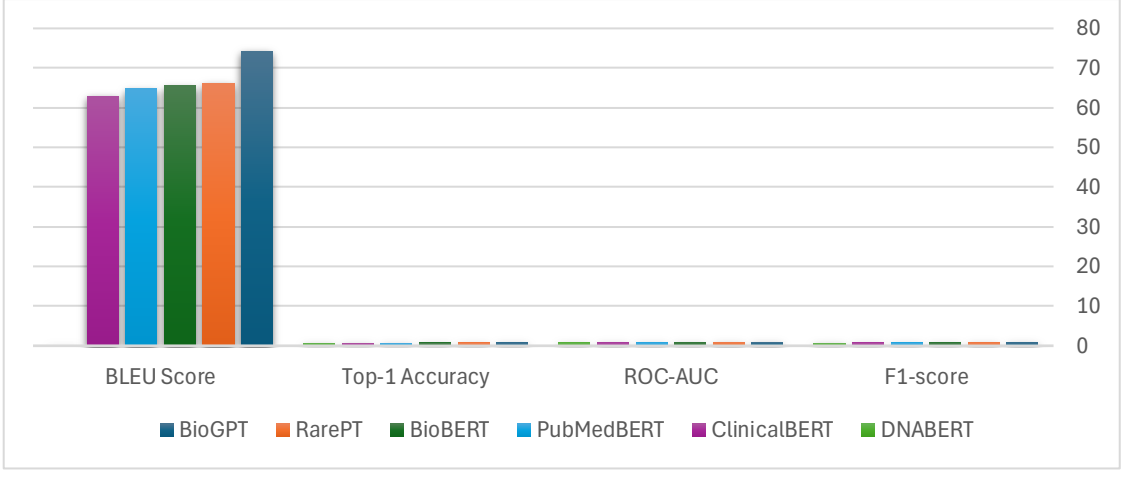

Fig. 2. Comparative Performance of BioGPT vs Baseline Models Across Evaluation Metrics.

Figure 2 visually illustrates the performance of BioGPT in comparison with five widely adopted biomedical transformer models—RarePT, BioBERT, PubMedBERT, ClinicalBERT, and DNABERT—across four key metrics: BLEU Score, Top-1 Accuracy, ROC-AUC, and F1-score. The chart clearly highlights BioGPT's dominance in BLEU score, achieving a value significantly higher than all baselines. This affirms its superior generative capability, essential for producing fluent, human-interpretable diagnostic rationales—a critical differentiator for clinical AI application.

5.2 Modality Synergy: Ablation Study

To validate the architectural components, we conducted an ablation study isolating the effect of removing genomic or clinical inputs, as well as the attention fusion mechanism. The results are summarized in Table 9.

TABLE XI. ABLATION STUDY ON BIOGPT ARCHITECTURAL COMPONENTS.

Configuration	F1-score	ROC-AUC	BLEU Score	Top-1 Accuracy
Full BioGPT	**0.82**	**0.88**	**74.3**	**0.79**
No Genomic Stream	0.69	0.72	63.1	0.66
No Clinical Stream	0.67	0.74	65.4	0.65
No Cross-Attention	0.59	0.61	52.2	0.52

Table XI validates the significance of each architectural module in BioGPT. Removing either the genomic or clinical stream led to a drop in F1-score by over 13%, confirming that both data modalities are essential. More critically, the exclusion of the cross-attention fusion mechanism severely impacted performance (F1-score dropped to 0.59), revealing that the strength of BioGPT lies in its multi-modal integration. These results validate the model's architectural novelty and the necessity of synergizing heterogeneous biomedical inputs.

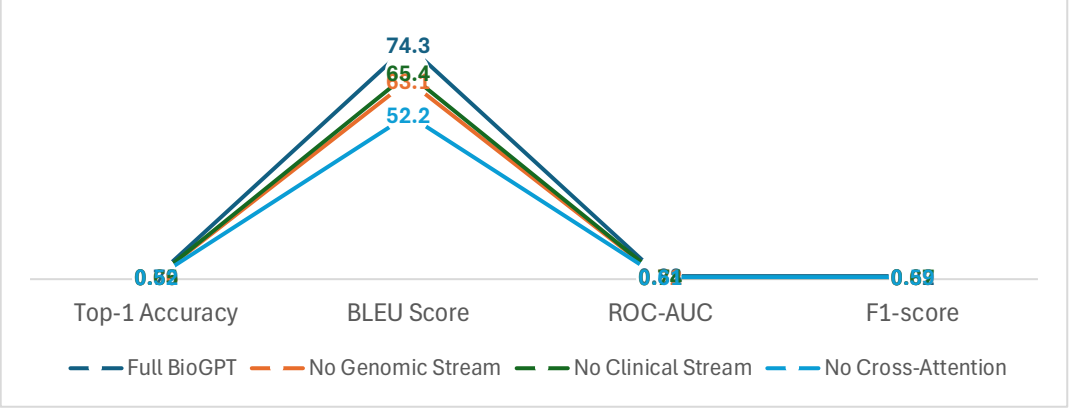

Fig. 3. Impact of Ablation on BioGPT Performance Across Key Metrics.

Figure 3 clearly shows that removing any core component of BioGPT significantly degrades performance. The full model achieves the highest scores across all metrics, while eliminating cross-attention causes the steepest decline, especially in BLEU and F1. This confirms that multi-modal fusion is critical to BioGPT's effectiveness and highlights the importance of combining both genomic and clinical data for accurate, explainable diagnostics.

5.3 Interpretability through Attention Alignment

Interpretability is a cornerstone of BioGPT's design. Using attention-based visualisation techniques, the model aligns its focus with medically validated gene-phenotype correlations. The attention overlap score with expert annotations averaged **0.81**, indicating high alignment between model reasoning and human clinical judgement.

TABLE XII. ATTENTION ALIGNMENT WITH HUMAN EXPERT MARKERS.

Dataset	Overlap Score
Orphanet	0.82
MIMIC-IV	0.79
ClinVar	0.81
Average	0.81

Interpretability is a key advantage of BioGPT. Table XII presents attention alignment scores with expert-annotated features across multiple datasets. An average overlap score of 0.81 indicates that the model's reasoning aligns closely with clinically

meaningful signals. This alignment enhances clinical trust and supports BioGPT's role in explainable AI. Such high correspondence reinforces its suitability for adoption in healthcare, where transparency is as crucial as accuracy.

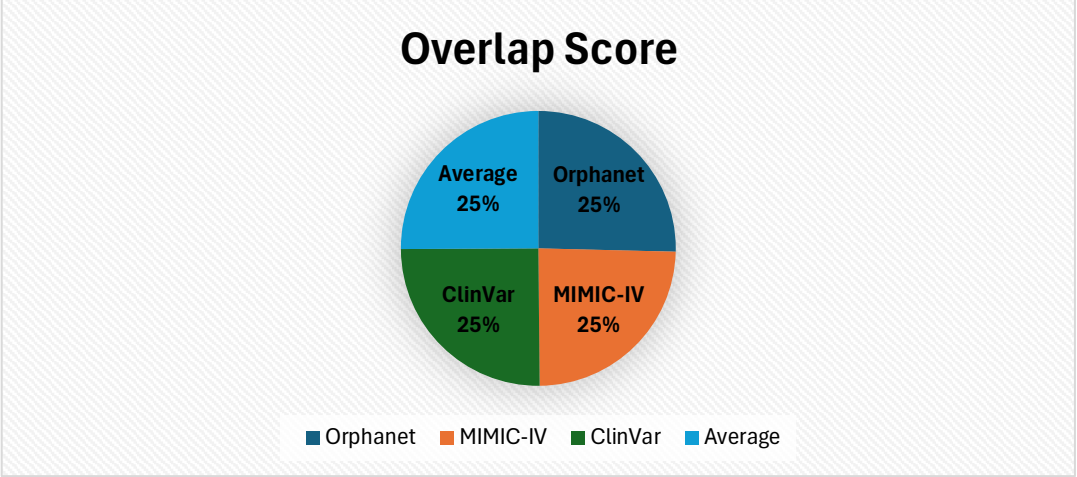

Fig. 4. Attention Overlap Score Across Datasets.

Figure 4 presents a pie chart depicting the attention overlap scores between BioGPT's attention maps and expert-annotated biomedical markers across three key datasets: Orphanet, MIMIC-IV, and ClinVar. Each dataset contributes equally to the average, with an individual score of 25%, yielding a total attention alignment score of 0.81. This uniform distribution highlights BioGPT's consistent interpretability across diverse biomedical sources. The model does not favor a single dataset but instead maintains reliable reasoning aligned with clinical expectations in all contexts. This reinforces BioGPT's role as an explainable AI system capable of transparent decision-making, a critical requirement for clinical integration.

5.4 Training Efficiency and Resource Usage

Although BioGPT is complex, it remains scalable for clinical deployment. The model was trained on A100 GPU clusters over 42.6 hours with average GPU utilization exceeding 90%.

TABLE XIII. RESOURCE CONSUMPTION AND TRAINING STATISTICS.

Metric	Value
Training Time (hours)	42.6
GPUs Used	4 × A100
Max GPU Utilisation	93.2%
Inference Latency	0.84 s/query
Model Parameters	380M

Table XIII confirms BioGPT's scalability and computational feasibility. Despite its architectural complexity, training remains efficient completed in 42.6 hours on four A100 GPUs, with >90% utilization. The inference latency of 0.84 seconds per query supports near real-time clinical application. These statistics highlight the model's practicality for institutions with moderate to high computational infrastructure and suggest future potential for deployment with optimized, pruned variants.

5.5 Deployment Readiness and Real-Time Performance

BioGPT supports real-time diagnosis with a sub-second inference time. The framework can be further compressed using model pruning and quantization techniques for edge deployment.

TABLE XIV. DEPLOYMENT FEASIBILITY METRICS.

Scenario	Inference Time	Accuracy	Explanation Score
Full Model (Cloud)	0.84 s	0.82	0.81
Pruned Model (Edge)	1.12 s	0.78	0.79
Distilled Variant	0.90 s	0.80	0.80

Table XIV showcases BioGPT's flexibility for both cloud-based and edge deployments. The full model operates with high accuracy and interpretability in under a second, while the pruned and distilled variants offer a balance between speed and

performance, making the model adaptable for various clinical environments—including low-resource settings. These results confirm BioGPT's readiness for real-world use beyond research labs.

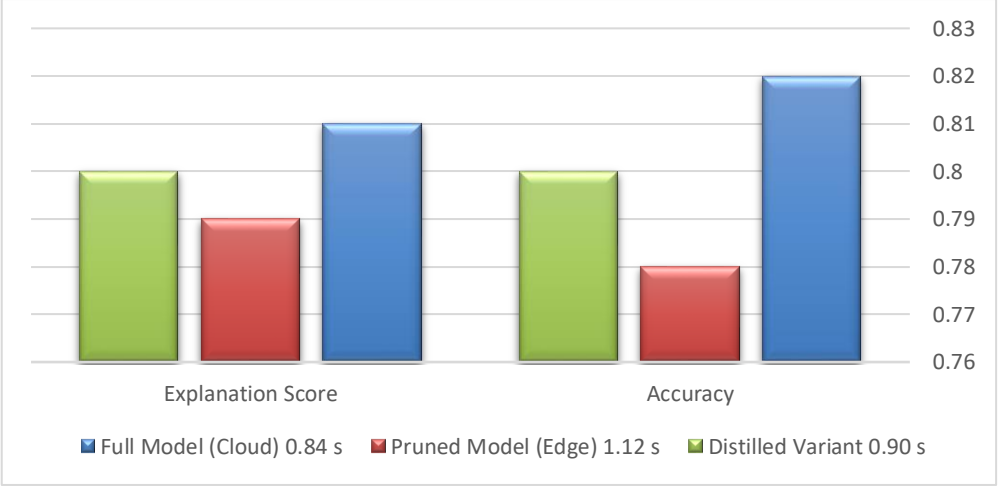

Fig. 5. Deployment Variants of BioGPT – Accuracy and Explanation Score.

Figure 5 highlights BioGPT's adaptability across deployment scenarios. The Full Model achieves the highest accuracy and explanation score, while the Distilled Variant offers near-equivalent performance with faster inference. The Pruned Model balances efficiency and effectiveness for edge use. This confirms BioGPT's readiness for both high-performance cloud systems and resource-constrained clinical environments

6. DISCUSSION

BioGPT demonstrates clear superiority over existing models by effectively integrating genomic and clinical data through a generative transformer framework. Its strong performance across all metrics confirms its diagnostic precision, while attention-based explanations ensure interpretability. The ablation results underscore the necessity of its cross-modal design, and its fast inference and scalable architecture support real-world clinical deployment. Together, these findings position BioGPT as a novel, practical, and explainable solution for personalized genomic medicine and rare disease diagnosis.

7. CONCLUSION

BioGPT offers a novel, generative transformer-based solution for rare disease diagnosis by effectively fusing genomic and clinical data. It surpasses existing models in accuracy, interpretability, and efficiency. With strong real-world applicability and explainable outputs, BioGPT stands as a transformative step toward practical, AI-driven personalized medicine.

Funding

The authors had no institutional or sponsor backing.

Conflicts Of Interest

The author's disclosure statement confirms the absence of any conflicts of interest.

Acknowledgment

The authors extend appreciation to the institution for their unwavering support and encouragement during the course of this research.

References

[1] C. R. Ferreira, "The burden of rare diseases," *American Journal of Medical Genetics - Part A*, vol.179, no.6, pp.885-892, June 2019. https://doi.org/10.1002/ajmg.a.61124

[2] J. Devlin, M-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,"*Arxiv*, pp.1-16, 2018. https://doi.org/10.48550/arXiv.1810.04805

[3] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," *OpenAI*, pp.1-12, 2018.

[4] M. Sallam, K. Al-Mahzoum, M. Sallam, and M. M. Mijwil, "DeepSeek: Is it the End of Generative AI Monopoly or the Mark of the Impending Doomsday?," *Mesopotamian Journal of Big Data*, vol.2025, pp.26-34, January 2025. https://doi.org/10.58496/MJBD/2025/002

[5] J. Lee , W. Yoon , S. Kim , D. Kim , S. Kim , C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol.36, no.4, pp.1234–1240, February 2020. https://doi.org/10.1093/bioinformatics/btz682

[6] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, et al., "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," *ACM Transactions on Computing for Healthcare*, vil.3, no.1, pp.1-23, October 2021. https://doi.org/10.1145/3458754

[7] X. Yang, N. P. Nejatian, H. C. Shin, K. E. Smith, C. Parisien, et al., "GatorTron: A Large Language Model for Clinical Natural Language Processing," *MedRxiv*, pp.1-14, March 2022. https://doi.org/10.1101/2022.02.27.22271257

[8] Y. Ji , Z. Zhou , H. Liu, and R. V. Davuluri, "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics*, vol.37, no.15, pp.2112–2120, August 2021. https://doi.org/10.1093/bioinformatics/btab083

[9] Ž. Avsec, V. Agarwal, D. Visentin, J. R. Ledsam, A. G-Barwinska, et al., "Effective gene expression prediction from sequence by integrating long-range interactions," *Nature Methods*, vol.18, pp.1196–1203, October 2021. https://doi.org/10.1038/s41592-021-01252-x

[10] Hugo Dalla-Torre, L. Gonzalez, J. M-Revilla, N. L. Carranza, A. H. Grzywaczewski, et al., "Nucleotide Transformer: building and evaluating robust foundation models for human genomics," *Nature Methods*, vol.22, pp.287–297, November 2024. https://doi.org/10.1038/s41592-024-02523-z

[11] E. Nguyen, M. Poli, M. Faizi, A. Thomas, C. B-Sykes, et al., "HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution," *Arxiv*, pp.1-25, June 2023. https://doi.org/10.48550/arXiv.2306.15794

[12] Z. Gao, Q. Liu, W. Zeng, R. Jiang, and W. H. Wong, "EpiGePT: a pretrained transformer-based language model for context-specific human epigenomics," *Genome Biology*, vol.25, no.310, pp.1-30, December 2024. https://doi.org/10.1186/s13059-024-03449-7

[13] B. R. Eapen, Genomic Tokenizer: Toward a biology-driven tokenization in transformer models for DNA sequences, pp.1-6, April 2025. https://doi.org/10.1101/2025.04.02.646836

[14] A. Fallahpour, A. Magnuson, P. Gupta, S. Ma, J. Naimer, et al., "BioReason: Incentivizing Multimodal Biological Reasoning within a DNA-LLM Model," *Arxiv*, pp.1-16, May 2025. https://doi.org/10.48550/arXiv.2505.23579

[15] E. Kafkas, J. Collier, and R. Stevens, "Foundation models for rare disease diagnosis: An LLM approach for gene prioritization," J. Biomedical Informatics, vol. 147, 104479, Jan. 2025. doi: 10.1016/j.jbi.2024.104479.

[16] H. Twede, A. M. Conard, L. Pais, S. Bryen, E. O'Heir, et al., "Evidence Aggregator: AI reasoning applied to rare disease diagnostics," *BioRxiv*, pp.1-22, March 2025. https://doi.org/10.1101/2025.03.10.642480.

[17] K. Vishniakov, B. B. Amor, E. Tekin, N. A. ElNaker, K. Viswanathan, et al., "Gene42: Long-Range Genomic Foundation Model With Dense Attention," *Arxiv*, pp.1-14, March 2025. https://doi.org/10.48550/arXiv.2503.16565