Can Large Language Models Classify and Generate Antimicrobial Resistance Genes?

Hyunwoo Yoo[1] **Haebin Shin**[2] **Gail Rosen**[1]

[1]Drexel University [2]KAIST AI
{hty23, glr26}@drexel.edu haebin.shin@kaist.ac.kr

Abstract

This study explores the application of generative Large Language Models (LLMs) in DNA sequence analysis, highlighting their advantages over encoder-based models like DNABERT2 and Nucleotide Transformer. While encoder models excel in classification, they struggle to integrate external textual information. In contrast, generative LLMs can incorporate domain knowledge, such as BLASTn annotations, to improve classification accuracy even without fine-tuning. We evaluate this capability on antimicrobial resistance (AMR) gene classification, comparing generative LLMs with encoder-based baselines. Results show that LLMs significantly enhance classification when supplemented with textual information. Additionally, we demonstrate their potential in DNA sequence generation, further expanding their applicability. Our findings suggest that LLMs offer a novel paradigm for integrating biological sequences with external knowledge, bridging gaps in traditional classification methods.

1 Introduction

Language Models (LMs) have demonstrated remarkable performance in various Natural Language Processing (NLP) tasks and have recently gained attention in bioinformatics, particularly in DNA sequence analysis. Encoder-based transformer models, such as DNABERT (Ji et al., 2021; Zhou et al., 2023), ProteinBERT (Brandes et al., 2022) and Nucleotide Transformer (Dalla-Torre et al., 2023), have shown strong performance in DNA sequence classification, leveraging nucleotide tokenization and self-supervised pretraining. These models are widely adopted for gene sequence analysis, promoter prediction, and mutation impact assessment. However, encoder-based models have inherent limitations in integrating external domain knowledge. Their fixed input structures make it difficult to incorporate additional textual context, such as BLASTn search results, which often contain critical biological insights. Additionally, these models may struggle to generalize when a single DNA sequence is associated with multiple labels, requiring strict pre-defined training paradigms.

Generative Large Language Models (LLMs), such as GPT-based models (Brown et al., 2020), introduce greater flexibility by allowing predictions to be influenced by external knowledge via prompting. Unlike encoder-based models, generative LLMs can dynamically incorporate supplementary textual information, which can enhance classification accuracy without requiring fine-tuning. Some biomedical LLMs, such as BioGPT (Luo et al., 2022) and Med-PaLM (Singhal et al., 2023), have demonstrated strong performance in processing medical and pharmaceutical text, but their applications in DNA sequence analysis remain underexplored. Beyond classification, generative LLMs also enable DNA sequence generation (Nguyen et al., 2024; Brixi et al., 2025), a capability that traditional encoder-based models lack. This opens new possibilities for exploring sequence design, mutation modeling, and synthetic data augmentation, further expanding the applicability of LLMs in genomics.

This study systematically evaluates the effectiveness of generative LLMs for DNA sequence classification and generation, comparing them against traditional encoder-based baselines. Our key contributions are as follows:

- We systematically compare generative LLMs with encoder-based models on DNA sequence classification tasks, providing a rigorous evaluation of their relative performance.
- We demonstrate that generative LLMs can leverage supplementary domain-specific knowledge to improve classification accuracy even without fine-tuning.

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 240–248
August 1, 2025 ©2025 Association for Computational Linguistics

- We explore the potential of generative LLMs in DNA sequence generation, analyzing their ability to generate biologically meaningful sequences and their implications for synthetic data augmentation.

Our findings suggest that generative LLMs offer a novel paradigm for integrating DNA sequences with external knowledge sources, expanding their applicability in bioinformatics research.

2 Related Works

Transformer-based encoder models have been widely applied to DNA sequence classification. DNABERT (Ji et al., 2021; Zhou et al., 2023) applies self-supervised learning to nucleotide sequences using k-mer tokenization, while DNABERT2 improves efficiency by introducing byte pair encoding (BPE) (Zhou et al., 2023). Nucleotide Transformer (Dalla-Torre et al., 2023) extends this approach by pretraining on diverse genomic datasets, achieving strong performance in gene classification tasks.

While these models perform well in classification, they have limited ability to incorporate external domain knowledge, such as BLASTn annotations (Lobo, 2008). Moreover, they struggle with handling multi-label classification, which is common in genomic studies (Bonin et al., 2023a; Marini et al., 2022). Our work differs by exploring whether generative LLMs can improve classification performance by dynamically integrating external textual information without additional fine-tuning.

Generative Large Language Models (LLMs) such as GPT-based models (Brown et al., 2020) have demonstrated strong natural language understanding but have been rarely applied to DNA sequence analysis. BioGPT (Luo et al., 2022), for example, is trained on biomedical literature but lacks direct training on DNA sequences.

Unlike encoder-based models, LLMs can dynamically incorporate supplementary textual information, such as BLASTn search results (Lobo, 2008), potentially enhancing classification performance. Additionally, LLMs have the potential for DNA sequence generation, which can be applied to mutation modeling and synthetic data augmentation, as demonstrated in previous studies exploring deep learning methods for genomic analysis (Marini et al., 2022; Arango-Argoty et al., 2018; Lakin et al., 2019).

While prior studies have focused on applying LLMs to biomedical text, our approach investigates whether generative LLMs can be effectively utilized for both classification and sequence generation in DNA analysis, providing a flexible alternative to traditional encoder-based models.

3 Methods

3.1 Data Collection

The dataset used in this study consists of antibiotic resistance gene sequences collected from the MEGARes (Doster et al., 2020; Bonin et al., 2023b) and CARD databases (Jia et al., 2017). The labels from MEGARes and CARD were mapped using the Antibiotic Resistance Ontology from the European Bioinformatics Institute (Cook et al., 2016), following previous research methods (Yoo et al., 2024). These databases contain DNA sequences associated with antimicrobial resistance (AMR) and provide multi-label annotations, where a single sequence may belong to multiple resistance categories. To incorporate external domain knowledge, we applied the BLASTn algorithm (Chen et al., 2015) to identify sequences similar to each DNA sequence in the dataset. For each sequence, the top-5 BLASTn search results were selected based on the e-value criterion, and their corresponding functional annotations were extracted. This additional textual information includes gene descriptions, known resistance mechanisms, and sequence alignment details, which were later integrated into our LLM-based classification prompts.

3.2 Baseline Models and Preprocessing

To compare the performance of generative LLMs with existing DNA sequence classification models, we included encoder-based models as baselines: DNABERT2, and Nucleotide Transformer. DNABERT2 (Zhou et al., 2023) is a BERT-based model to process DNA sequences as natural language text. It improved version of DNABERT (Ji et al., 2021) by introducing byte pair encoding (BPE) instead of utilizing k-mer tokenization, allowing for more efficient sequence representation. Nucleotide Transformer (Dalla-Torre et al., 2023), a transformer model pre-trained on diverse genomic datasets, has demonstrated strong performance in various molecular phenotype prediction tasks.

For all models, DNA sequences were preprocessed by converting them to uppercase, and in-

Model	Accuracy	Precision	Recall	F1 Score
DNABERT2 (Finetuning)	0.8697	0.8161	0.6996	0.7332
Nucleotide Transformer (Finetuning)	0.8289	0.8184	0.5867	0.6579
LLama3.1 8B-4bit	0.0037	0.0011	0.0002	0.0003
LLama3.1 8B-4bit + Blastn	0.0744	0.0530	0.0129	0.0207
LLama3.1 8B-4bit + Finetuning	0.5521	0.4760	0.5521	0.5080
Claude3.5sonet	0.1488	0.1770	0.0966	0.0735
Claude3.5sonet + Blastn	0.8042	0.6287	0.5421	0.5794
Chatgpt4o-mini	0.00	0.00	0.00	0.00
Chatgpt4o-mini + Blastn	0.7804	0.9090	0.7804	0.8398
Chatgpt4o-mini + Finetuning	0.9318	0.9337	0.9318	0.9319

Table 1: Performance metrics for DNA sequence classification across multiple models. Chatgpt4o-mini with finetuning achieves the best overall performance, with the highest accuracy and F1 score, surpassing specialized models like DNABERT2 and the Nucleotide Transformer. Without finetuning or external features, general-purpose LLMs such as LLaMA3.1, Claude3.5, and Chatgpt4o-mini perform poorly, indicating that both biological context via BLASTn and domain-specific adaptation are critical for DNA sequence understanding.

valid sequences were removed. The final dataset consisted only of validated antibiotic resistance gene sequences.

3.3 Fine-tuning and Prompt-based Classification

To evaluate generative LLMs in DNA sequence classification, we employed Low-Rank Adaptation (LoRA) fine-tuning on LLaMA. LoRA enables parameter-efficient tuning by modifying only a subset of the model's weight matrices, significantly reducing computational costs while maintaining performance (Hu et al., 2021). Further details provided in Appendix B. Additionally, we conducted zero-shot inference using the Claude 3.5 sonet (Anthropic, 2024) and ChatGPT-4 API to assess how well proprietary generative models classify DNA sequences without explicit training. To investigate whether generative LLMs can classify DNA sequences without fine-tuning, we formulated two experimental settings. In the first setting, models were given only the DNA sequence as input. In the second setting, models received both the DNA sequence and the top-5 BLASTn search results, including functional annotations and gene descriptions. This setup allowed us to assess whether LLMs can leverage external domain knowledge to improve classification accuracy. Prompt details are in Appendix A

3.4 LLM-Based DNA Sequence Generation

In addition to classification, we explored whether generative LLMs can synthesize biologically meaningful DNA sequences. To this end, we designed a sequence generation task where models were prompted with initial part of antibiotic resistance genes and tasked with generating plausible full DNA sequences. We implemented finetuning with LLMs. Further details provided in Appendix C. The quality of the generated sequences was assessed using three similarity measures. Levenshtein (Levenshtein, 1966) distance was used to quantify the sequence-level similarity by measuring the minimum number of edits (insertions, deletions, and substitutions) required to match a reference sequence. Jaccard's index of similarity (Real and Vargas, 1996) was computed on k-mer tokenized sequences to evaluate overlapping subsequences between generated and known resistance gene sequences. Cosine similarity was also applied to k-mer frequency vectors to compare overall sequence composition (Ng, 2017). To ensure basic functional validity, we checked whether generated sequences maintained proper nucleotide composition. GC content (Marmur and Doty, 1962) distribution was compared to existing AMR sequences to verify biological plausibility.

4 Experiments

4.1 DNA Classification

To evaluate the performance of generative language models in DNA sequence classification, we conducted experiments under three conditions. The first setting involved using the base models, where only the raw DNA sequence was provided as input. The second setting introduced BLASTn search results, incorporating additional textual annotations such as gene descriptions and resistance mechanisms. The third setting involved fine-tuning the models using labeled DNA sequences. For baseline comparisons, we included DNABERT2 and Nucleotide Transformer, which have demonstrated strong performance in DNA sequence classification tasks. The generative models evaluated in this study include LLaMA 3.1 (Meta AI, 2024) (8B-4bit), Claude 3.5 Sonet, and ChatGPT-4o-mini. Each model was tested in zero-shot, BLASTn-

Model	Jaccard Similarity	Cosine Similarity	Levenshtein Similarity	GC Correlation
GENERater (Zero-shot)	0.9970	0.9680	0.3790	0.8436
ChatGPT-4o-mini (Fine-tuned API)	0.9870	0.9857	0.5776	0.7930
GENErator (LoRA Fine-tuned)	0.9970	0.9680	0.3790	0.8436
LLaMA 3.2 1B (LoRA Fine-tuned)	0.2659	0.5911	0.2004	0.6938
Gemma 3 1B (LoRA Fine-tuned)	0.3177	0.7580	0.2487	0.7305

Table 2: Similarity scores between generated DNA sequences and the ground truth across various models. GENERater, both in zero-shot and LoRA fine-tuned settings, achieves near-perfect Jaccard and Cosine similarities, with strong GC content correlation, indicating high biological fidelity. ChatGPT-4o-mini also performs competitively despite being a general-purpose LLM. In contrast, smaller fine-tuned models like LLaMA 3.2 1B and Gemma 3 1B yield significantly lower similarity scores across all metrics, highlighting the challenge of DNA generation in low-resource model settings.

augmented, and fine-tuned configurations.

4.2 DNA Sequence Generation

In addition to classification, we assessed whether generative language models could synthesize biologically meaningful DNA sequences. A dataset of antimicrobial resistance genes from Acinetobacter baumannii was collected using the NCBI Entrez API, with 1,000 sequences retrieved. The dataset was split into 80% for training and 20% for testing. Input sequences were trimmed to a length of 200 base pairs, while the maximum generated output length was set to 3,000 base pairs. Further details on dataset characteristics provided in Appendix E. For baseline comparisons, we included GENERater (Wu et al., 2025), which were evaluated in a zero-shot setting. For fine-tuned models, we used ChatGPT-4o-mini finetuned via API along with GENErator, LLaMA 3.2 1B (Grattafiori et al., 2024), and Gemma 3 1B (Gemma Team, Google DeepMind, 2025), which were finetuned using the LoRA. Each model was assessed based on its ability to generate sequences that resemble known antimicrobial resistance genes.

5 Results and Discussion

Table 1 presents the classification results across various model configurations. Encoder-based models, DNABERT2 and Nucleotide Transformer, consistently demonstrated the highest accuracy, with DNABERT2 achieving 86.97% accuracy and Nucleotide Transformer reaching 82.89%. In contrast, generative models performed poorly in the base setting, with LLaMA 3.1 obtaining an accuracy of only 0.37%. Considering this outcome alongside the unclassified rate reported in Appendix D, it appears that generative models have difficulty performing direct DNA sequence classification without supplementary context. The inclusion of BLASTn search results significantly improved classification accuracy. ChatGPT-4o-mini, which initially failed to classify any sequences correctly, achieved 78.04% accuracy with BLASTn augmentation. Similarly, Claude 3.5 Sonet improved from 14.88% to 80.42% accuracy. These results suggest that LLMs benefit from external textual information, compensating for their lack of prior exposure to DNA sequences. Fine-tuning further enhanced classification accuracy, with ChatGPT-4o-mini achieving 93.18%, surpassing DNABERT2. This demonstrates that while LLMs struggle in a zero-shot setting, targeted training on DNA sequences allows them to match or exceed the performance of specialized encoder-based models.

Table 2 summarizes the similarity scores for generated DNA sequences. In the zero-shot setting, GENERater produced sequences with high Jaccard similarity (0.9970) and Cosine similarity (0.9680), but relatively low Levenshtein similarity (0.3790), indicating that while generated sequences share common k-mers with known resistance genes, their exact sequence composition differs significantly. Fine-tuned models exhibited varying levels of similarity. ChatGPT-4o-mini, fine-tuned via API, achieved the highest similarity across all three metrics, particularly in Levenshtein similarity (0.5776), suggesting that it generated sequences more closely aligned with known resistance genes at the character level. GENErator (LoRA Fine-tuned) maintained nearly identical similarity scores to its zero-shot counterpart, whereas LLaMA 3.2 1B and Gemma 3 1B displayed substantially lower similarity scores across all metrics, indicating challenges in generating sequences that closely resemble existing DNA. Further analysis of GC content confirmed that fine-tuned models generated biologically plausible sequences. However, additional validation is required to determine whether these sequences retain functional properties relevant to antimicrobial resistance.

6 Conclusion

This study demonstrated that generative LLMs offer greater flexibility in DNA sequence classification and generation compared to traditional encoder-based models. While encoder models like DNABERT2 performed well in standard classification tasks, generative models benefited significantly from additional textual information, highlighting their ability to integrate external domain knowledge. Fine-tuned generative models also produced biologically plausible DNA sequences, suggesting potential applications in synthetic biology. However, LLMs struggled in zero-shot classification, emphasizing the need for fine-tuning and improved biological data integration.

7 Limitations

While this study highlights the potential of generative LLMs in DNA sequence analysis, there are several areas for further improvement. Zero-shot classification performance remained limited, underscoring the need for fine-tuning or integrating external biological knowledge to enhance prediction accuracy. Future work could explore hybrid approaches that combine LLMs with domain-specific models or structured databases to improve robustness.

In DNA sequence generation, fine-tuned models successfully produced sequences structurally similar to known antimicrobial resistance genes. However, additional real-world validation through laboratory experiments is necessary to determine whether these sequences retain functional properties relevant to resistance mechanisms.

Another key consideration is the computational cost associated with fine-tuning large-scale models. The substantial resource requirements highlight the need for more efficient adaptation techniques, such as parameter-efficient fine-tuning or retrieval-augmented approaches. Future research should investigate methods to balance computational efficiency with model performance to enable broader accessibility and practical applications in bioinformatics.

Acknowledgments

This work was supported in part by the National Science Foundation (NSF) under Grant Number 2107108.

References

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic. https://www.anthropic.com/index/claude-3-opus-sonnet-haiku.

Gustavo Arango-Argoty, Emily Garner, Amy Pruden, Lenwood S. Heath, Peter Vikesland, and Liqing Zhang. 2018. Deeparg: A deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6(1):23.

Nathalie Bonin, Enrique Doster, Hannah Worley, Lee J Pinnell, Jonathan E Bravo, Peter Ferm, Simone Marini, Mattia Prosperi, Noelle Noyes, Paul S Morley, and Christina Boucher. 2023a. Megares and amr++, v3.0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. *Nucleic Acids Research*, 51(D1):D744–D752.

Nathalie Bonin, Enrique Doster, Hannah Worley, Lee J Pinnell, Jonathan E Bravo, Peter Ferm, Simone Marini, Mattia Prosperi, Noelle Noyes, Paul S Morley, and Christina Boucher. 2023b. MEGARes and AMR++, v3.0: an updated comprehensive database of antimicrobial resistance determinants and an improved software pipeline for classification using high-throughput sequencing. *Nucleic Acids Research*, 51(D1):D744–D752.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110.

Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, Jonathan C. Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. 2025. Genome modeling and design across all domains of life with evo 2. *bioRxiv*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric

Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Ying Chen, Weicai Ye, Yongdong Zhang, and Yuesheng Xu. 2015. High speed blastn: an accelerated megablast search tool. *Nucleic Acids Research*, 43(16):7762–7768.

Charles E. Cook, Mary Todd Bergman, Robert D. Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler. 2016. The european bioinformatics institute in 2016: Data growth and integration. *Nucleic Acids Research*, 44(D1):D20–D26.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, et al. 2023. The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *Genomics*.

Enrique Doster, Steven M Lakin, Christopher J Dean, Cory Wolfe, Jared G Young, Christina Boucher, Keith E Belk, Noelle R Noyes, and Paul S Morley. 2020. Megares 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Research*, 48(D1):D561–D569.

Gemma Team, Google DeepMind. 2025. Gemma 3 technical report. Technical report, Google DeepMind. See Contributions and Acknowledgments section for full author list.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv*. ArXiv:2106.09685v2.

Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120.

Baofeng Jia, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, Biren M. Dave, Sheldon Pereira, Arjun N. Sharma, Sachin Doshi, Mélanie Courtot, Raymond Lo, Laura E. Williams, Jonathan G. Frye, Tariq Elsayegh, Daim Sardar, Erin L. Westman, Andrew C. Pawlowski, Timothy A. Johnson, Fiona S.L. Brinkman, Gerard D. Wright, and Andrew G. McArthur. 2017. Card 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45(D1):D566–D573.

Steven M. Lakin, Alan Kuhnle, Bahar Alipanahi, Noelle R. Noyes, Chris Dean, Martin Muggli, Rob Raymond, et al. 2019. Hierarchical hidden markov models enable accurate and diverse detection of antimicrobial resistance sequences. *Communications Biology*, 2(1):294.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics-Doklady*, 10(8):707–710. Translated from Doklady Akademii Nauk SSSR, Vol. 163, No. 4, pp. 845-848, August 1965.

Ingrid Lobo. 2008. Basic local alignment search tool (blast). *Nature Education*, 1(1):215. © 2008 Nature Education.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Simone Marini, Marco Oliva, Ilya B Slizovskiy, Rishabh A Das, Noelle Robertson Noyes, Tamer Kahveci, Christina Boucher, and Mattia Prosperi. 2022. Amr-meta: A k -mer and metafeature approach to classify antimicrobial resistance from high-throughput short-read metagenomics data. *GigaScience*, 11. Giac029.

J. Marmur and P. Doty. 1962. Determination of the base composition of deoxyribonucleic acid from its thermal denaturation temperature. *Journal of Molecular Biology*, 5:109–118.

Meta AI. 2024. Introducing llama 3.1: Our most capable models to date. `https://ai.meta.com/blog/llama-3-1`. Accessed: 2025-03-20.

Patrick Ng. 2017. dna2vec: Consistent vector representations of variable-length k-mers. *arXiv preprint arXiv:1701.06279*.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. 2024. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723):eado9336.

Raimundo Real and Juan M. Vargas. 1996. The probabilistic basis of jaccard's index of similarity. *Systematic Biology*, 45(3):380–385.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620:172–180. Publisher Correction published on 27 July 2023.

Wei Wu, Qiuyi Li, Mingyang Li, Kun Fu, Fuli Feng, Jieping Ye, Hui Xiong, and Zheng Wang. 2025. GENERator: A Long-Context Generative Genomic Foundation Model. *arXiv preprint*, arXiv:2502.07272.

Hyunwoo Yoo, Bahrad Sokhansanj, James R. Brown, and Gail Rosen. 2024. Predicting anti-microbial resistance using large language models. *arXiv preprint arXiv:2401.00642*. `https://doi.org/10.48550/arXiv.2401.00642`.

Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. 2023. Dnabert-2: Efficient foundation model and benchmark for multi-species genome. *arXiv*. ArXiv:2306.15006v1.

A Example Prompts

A.1 Example Prompt Explanation including DNA Sequence

In this example prompt, a DNA sequence is provided along with several drug class labels, such as Sulfonamides, Aminoglycosides, Betalactams, Glycopeptides, Tetracyclines, Phenicol, Fluoroquinolones, MLS (Macrolide-Lincosamide-Streptogramin), and Multi-drug resistance. The task involves asking the model to determine the drug class that the DNA sequence is resistant to.

The prompt follows this format:

```
"Tell me the resistance drug among drugs (Sulfonamides, Aminoglycosides, Betalactams, Glycopeptides, Tetracyclines, Phenicol, Fluoroquinolones, MLS, Multi-drug_resistance) with DNA sequence (ATGAATCCCTATC... ...ACAAACTGCGAGGCAGTTCGCATGA)?"
```

This prompt is used to assess the DNA sequence for antibiotic resistance and classify the sequence into one of the specified drug resistance categories.

A.2 Example Prompt Explanation including Blastn information

In this prompt, a DNA sequence and the top 5 Blastn search results are provided. The task is to predict the drug class that the DNA sequence is resistant to, based on the alignment information and matching sequences. The drug class labels included in the prompt are Sulfonamides, Aminoglycosides, Betalactams, Glycopeptides, Tetracyclines, Phenicol, Fluoroquinolones, MLS (Macrolide-Lincosamide-Streptogramin), and Multi-drug resistance.

The BLASTn results contain gene information such as sequence titles, alignment length, e-values, and detailed sequence alignments (query, match, and subject sequences). This allows the model to analyze the DNA sequence's pattern and classify it into the appropriate drug resistance category.

The prompt follows this format:

```
"Tell me the resistance drug among drugs (Sulfonamides, Aminoglycosides, Betalactams, Glycopeptides, Tetracyclines, Phenicol, Fluoroquinolones, MLS, Multi-drug_resistance) with DNA information ([{'sequence_title': 'gi|1035502645|ref|NG_048504.1| Enterococcus casseliflavus vanXY-C gene for D-Ala-D-Ala dipeptidase/D-Ala-D-Ala carboxypeptidase VanXY-C, complete CDS', 'alignment_length': 673, 'e_value': 0.0, 'query_sequence': 'ATGAATCCCTATCTA...', 'match_sequence': '||||||||||||||...', 'subject_sequence': ...'}, ... ])?"
```

This prompt aims to predict the antibiotic resistance drug by using DNA sequence data from the Blastn search results and identifying the relevant drug resistance class.

B Finetuning of DNA Sequence Classification Models

The Meta-LLaMA-3.1-8B-Instruct model was fine-tuned using the Unsloth framework with 4-bit quantization to enhance memory efficiency. A LoRA configuration was applied to key projection layers, with moderate values for the rank and scaling parameters. The training dataset consisted of DNA sequences and their associated resistant drug class labels, organized in a system-user-assistant conversational format and later converted to the Alpaca-style instruction-following format. Each example included instruction, input, and output fields, and samples were padded with an end-of-sequence token. Training was conducted using the SFTTrainer with mixed-precision enabled (fp16 or bf16), depending on hardware availability. In addition, a GPT-based model (gpt-4o-mini-2024-07-18) was customized using task-specific instruction examples via the OpenAI fine-tuning API.

C Finetuning of DNA Sequence Generation Models

Three large language models (LLMs) were fine-tuned for DNA sequence generation using parameter-efficient fine-tuning (PEFT) with LoRA. The dataset contained DNA input-output sequence pairs, tokenized using model-specific tokenizers and padded using the end-of-sequence token. LoRA configurations were adjusted for each model, with common values for rank, scaling, and dropout, and target modules selected based on the architecture. Training was conducted for several epochs with standard optimization settings.

The GENERator-eukaryote-3b-base model used separate tokenization strategies for inputs and outputs, with padding tokens in the labels replaced by -100. LoRA was applied to selected attention projections, and training used fp16 precision. The Llama-3.2-1B model supported sequences up to 4096 tokens and followed a prompt format of "Input: <input_sequence> Output: <output_sequence>", using bf16 precision and a memory-efficient optimizer. The Gemma-3-1B-PT model followed a similar formatting and applied LoRA to a subset of projection layers, using float32 precision to ensure stability. A GPT-based model (gpt-4o-mini-2024-07-18) was additionally adapted through OpenAI's fine-tuning API using domain-specific examples.

D Unclassified Rate

Additional gene information from the Blastn DB search results was provided, performance improved even without additional training on this data. As seen in Table 3, the Unclassified Rate decreased across all models. For the LLaMA 3.1 8B-4bit quantized model, the rate dropped from 97% to 73% when using Blastn. For Claude 3.5 sonet, it decreased from 39% to 11%. ChatGPT 4-mini showed a sharp improvement, going from classifying nothing to only leaving 14% unclassified. When fine-tuning was applied, both the LLaMA 3.1 8B 4bit quantized model and ChatGPT 4-mini reduced their unclassified rates to 0%.

Model	Unclassified Rate
LLama3.1 8B-4bit (Base Model)	97%
LLama3.1 8B-4bit (Blastn)	73%
LLama3.1 8B-4bit (Finetuning)	0%
Claude3.5sonet (Base Model)	39%
Claude3.5sonet (Blastn)	11%
Chatgpt4o-mini (Base Model)	100%
Chatgpt4o-mini (Blastn)	14%
Chatgpt4o-mini (Finetuning)	0%

Table 3: Model unclassified rates with long names displayed in two lines.

E DNA Generation Dataset Characteristics

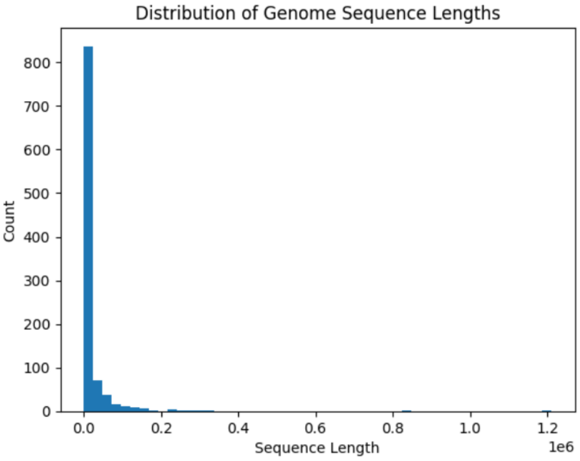

Figure 1: Length distribution of the Acinetobacter baumannii dataset.

Acinetobacter baumannii is a Gram-negative, opportunistic pathogen that poses a serious threat in healthcare settings due to its ability to survive in diverse environments and its remarkable capacity for antibiotic resistance. This bacterium is known for forming robust biofilms, which enhance its persistence on medical equipment and hospital surfaces. It exhibits resistance to multiple antibiotic classes, including beta-lactams, aminoglycosides, and fluoroquinolones, primarily through mechanisms such as enzymatic degradation (e.g., beta-lactamases), efflux pumps, and target site modifications. Given its clinical significance and growing prevalence in multidrug-resistant infections, we collected 1,000 sequences of Acinetobacter baumannii using the NCBI Entrez API for further analysis.

The dataset characteristics are summarized below:

Sequence Statistic	**Length (bp)**
Average sequence length	16,325.75
Median sequence length	1,033.50
Minimum sequence length	204
Maximum sequence length	1,210,760

Table 4: Statistics of the collected Acinetobacter baumannii sequences

The length distribution of the dataset is shown in Figure 1. The length distribution of the dataset exhibits a wide range, spanning from 204 bp to over 1.2 million bp, with a median length of approximately 1,033.50 bp. The substantial difference between the median and the mean (16,325.75 bp) suggests a right-skewed distribution, indicating the presence of a small number of extremely long sequences. Such distribution may impact downstream analysis, particularly in tasks such as sequence alignment or model training, where extreme sequence lengths might introduce computational challenges.

Additionally, the presence of very short sequences (minimum: 204 bp) suggests that preprocessing steps such as length filtering or normalization may be necessary to ensure consistency in downstream analyses. A closer examination of the length distribution (as illustrated in Figure 1) could provide further insights into potential clustering patterns or the need for stratified handling of different length groups.