IJARSCT

International Journal of Advanced Research in Science, Communication and Technology

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

ISSN: 2581-9429

Volume 6, Issue 1, January 2026

Impact Factor: 7.67

Integrating DNA Cryptography and Large Language Models for Next-Generation Secure Communication

Anvitha J Shetty, Kavya R, Shreyas S Shetty, Srujan S, Dr. Kiran Y C
Information Science and Engineering,
Global Academy of Technology, Bengaluru, India

Abstract: *The DNA cryptography stores data in DNA's big space, many parts in parallel, and variability inside DNA parts, that makes it very secure. Large Language Models are good at creating and finding patterns in data, which constitutes their smartness and ability for continuous improvement overtime. Putting these two tools together solves problems in how to make keys, fix errors, and stay strong against new threats. Genome LM uses tokenization and transformer architecture for DNA encoding. Expected benefits are the increased key entropy, increased brute-force and statistical attack resistance, and improved error correction during decryption. But the impact of this discovery goes further than. First, this single use case yields revolutionary results for cloud computing. Medicine and finance alike, because it allows one to create a quick future- proof and scalable encryption method that is required for the constrained in resources; Internet of things IoT and edge devices.*

Keywords: Blockchain, Proof of Work, Consensus Algorithm, Cryptocurrency, Mining, Security, Distributed Systems, Byzantine Fault Tolerance system

I. INTRODUCTION

Data security is a big concern in this world today because we have data all over the place from cloud data to health records, banking systems, IoT devices and the list goes on. The cloud contains archives of billions of people and companies, of the health care systems with most sensitive data of our patients, of the Internet of Things, which are constantly generating personal data. While at the same time, cyberattacks are growing and the threat environment is evolving, quantum computing poses a potentially existential threat to encryption. These challenges need future generation cryptographic solutions that are not only secure but are also scalable, adaptable, and able to handle new and changing cyberattacks.

Traditional encryption algorithms such as RSA, AES, and ECC are based on complex mathematical problems. However, these may not be resistant to the power of a future quantum computer. In addition, previous systems have various weak points in the way keys are managed, data is stored, and difficult attacks are prevented. Moreover, they are not flexible enough to catch up and thus are less effective against new AI-driven cyberattacks.

DNA cryptography provides a promising new approach that takes advantage of DNA's huge data capacity, natural parallel processing, built-in randomness, and biological complexity. This encoding of plain text data into DNA sequences results in unpredictable and highly complex patterns that are difficult for attackers to decode. Thus, DNA encryption provides further security outside of digital encryption. Large Language Models have demonstrated impressive abilities in generating sequences, recognizing patterns, and learning adaptively has made them ideal for analyzing and optimizing sequence-based data like DNA.

LLMs are increasingly being used for the tasks of cryptanalysis, key generation, and enhancing encryption. algorithms, while in bioinformatics they showed a great potential for analyzing DNA and protein sequences. However, all the current DNA cryptography schemes suffer from crucial issues in key generation efficiency, error correction, and scalability. Although LLMs hold tremendous potential in sequence-based domains, they have

DOI: 10.48175/IJARSCT-30738

not been systematically used to enhance DNA cryptographic frameworks. This gap inspires the development of a hybrid model that brings together the unique strengths of DNA cryptography and LLMs.

This gap inspires the development of a hybrid model that brings together the unique strengths of DNA cryptography and LLMs. This paper presents a novel hybrid cryptographic frame- work that combines DNA encoding schemes with the LLM- guided key generation and error correction. The approach shows how LLMs can enhance the randomness, adaptability, and robustness of DNA-based encryption. The framework is evaluated against traditional DNA-only cryptography methods in terms of security, scalability, and decryption accuracy, highlighting its potential as a future generation solution for secure and safe communication.

DNA cryptography [1] [2] continues to be an evolving field that combines biological characteristics with computer science and cryptographic principals to achieve secure information encryption using the properties of DNA molec- ules. DNA is the basic life element which uses sequences of four nucleotides, which consist of adenine (A), cytosine (C), guanine (G) and thymine (T) components to form binary data sequences. The unique ability to encode large amount of information makes DNA as an exceptional storage device. By combining DNA sequences with standard encryption practices DNA cryptography establishes a fresh method to build se- cure communication channels. Security through unauthorized access gets enhanced by DNA's large data density as well as its complicated genetic pattern structure. The cryptographic methods using DNA can be used for encrypting data in ways that are both highly secure and resilient to many conventional cryptographic attacks. Tokenization in NLP is the process of breaking down a text into smaller units called tokens, which can be words, subwords, or characters (shown in Figure-1). The process of text data preparation requires tokenization as an essential step because it transforms unstructured text into a form suitable for NLP modeling algorithms to utilize. Through one-hot encoding NLP transforms categorical information including text corpus words into machine learning accessible numerical values. The NLP model uses one-hot encoding as a method to convert each unique word to a binary vector representation (Table-I). The binary vector contains a length matching the vocabulary size where it has one active "hot" position and all remaining positions remain at zero. Through its encoding method each word becomes distinct so that no ordinal connection exists between words. The fixed- length input requirement of models makes one-hot encoding highly useful because it enables word representation through binary vectors regardless of word frequency or contextual relevance.

II. BACKGROUND AND RELATED WORK

DNA cryptography utilizes the unique properties of DNA molecules and their nucleotide sequences (adenine [A], thymine [T], cytosine [C], and guanine [G]) to securely encode, store, and deliver data. Common encoding techniques include binary-to-
DNA mapping, where binary pairs
are translated into nucleotide bases (e.g., 00→A, 01→C, 10→G, 11→T), the use of DNA complementary rules based on Watson–Crick base pairing for encryption and decryption,
DNA-based XOR operations on sequences, and hybrid
methods that combine DNA encoding with traditional cryptographic algorithms such as AES and RSA to enhance security. Applications of DNA cryptography includes, secure digital data storage, steganography (placing hidden messages within DNA sequences), protection of sensitive medical and genomic data, and encryption systems for cloud computing and Internet of Things (IoT) environments.

Large Language Models (LLMs): Overview Large Language Models are deep neural networks trained on massive datasets to understand and generate complex human language as well as structured sequences. Most LLMs are built on the Transformer architecture, which uses self-attention mechanisms to capture long-range dependencies in data far more effectively than earlier models such as recurrent neural networks (RNNs) and long short- term memory (LSTM) networks. Their strengths include learning complex sequence patterns, generating structured outputs in different domains including DNA and RNA sequences, and applicability in cryptography for key generation, cryptanalysis, and optimization of encryption protocols. Within bioinformatics, LLMs have been increasingly used for tasks such as DNA and protein sequence prediction, mutation detection, and also modeling of biological data.

DOI: 10.48175/IJARSCT-30738

Relevant Advancements and Research Context- Recent DNA cryptography research highlights that DNA's massive parallelism and natural unpredictability, together offer resistance against brute-force attacks. Researchers are also exploring hybrid security models that combine DNA-based methods with classical encryption or neural network techniques to secure cloud and IoT platforms. At the same time, benchmarks like CipherBank and AICrypto show the increasing importance of LLMs in cryptographic reasoning, demonstrating clear advantages over older neural models in analysing biological sequences. Despite these developments, the use of modern LLM architectures in DNA cryptography is still limited. This gap presents an opportunity to design improved hybrid frameworks that bring together the strengths of biologically inspired DNA encryption and the adaptive and pattern recognition abilities of advanced LLMs.

Algorithm 1: How DNA Cryptography works:

```
Inputs:
Plaintext message p
A pre-trained LLM (for example, a Transformer-based model trained on genomic or text data
Binary-to-DNA encoding map M (e.g., 00→A, 01→C, 10→G, 11→T)
Outputs:
Encrypted DNA sequence E
A pre-trained LLM (for example, a Transformer-based model trained on genomic or text data
Binary-to-DNA encoding map M (e.g., 00→A, 01→C, 10→G, 11→T)
Encryptions Steps:
1:  Tokenize plaintext
2:  Binary conversion
3: Binary-to-DNA mapping
4: Generate pseudo-key
5: LLM key generation
6:  Encrypt message
7:  Transmit
Decryption Steps:
1:  Receive
2:  Regenerate key
3:  Decrypt
4. Error correction using LLM-based techniques
5. Recover plain text
```

III. RELATED-WORK

Recent studies have explored the intersection of DNA cryptography and large language models (LLMs), particularly transformer-based architectures, which have shown remarkable abilities for enhancing cryptographic schemes and modeling biological sequences and enhancing cryptographic structures.

Transformer-based genomic language models like DNABERT, Nucleotide, Transformer, and GenSLM have been developed to capture complex dependencies in DNA sequences with byte-pair encoding or tokenizing them into k-mers.In tasks like sequence classification, prediction, and representation learning, these models perform better and more then recurrent neural networks. Their good results indicates their potential for the application in secure DNA-based methods of encryption.

LLMs have been proved for key generation, cryptanalysis, and cryptographic protocol optimization. Benchmarks like CipherBank and AICrypto demonstrate that modern transformer LLMs have the capability of carrying out nontrivial cryptanalysis and key reasoning tasks. This indicates that enhanced by the LLM DNA cryptography systems new attack surfaces and opportunities.

For encoding, error correction, or key generation, existing DNA cryptography frameworks have mostly used classical neural networks like LLMs and CNNs; however, these methods do not fully utilize transformer LLMs' improved context

DOI: 10.48175/IJARSCT-30738

modeling and sequence generation capabilities. According to recent studies, integrating LLMs as intelligent agents that produce strong cryptographic keys from DNA sequences and enhance error correction during decryption may be possible.

To dynamically improve cryptographic strength, one innovative method combines DNA encoding schemes with LLM-driven secure key generation. According to Wang et al. (2024), this hybrid approach shows enhanced security in resource-constrained environments like IoT or cloud systems by periodically updating the LLM-based key generator to increase resilience against adversarial attacks.

The systematic application of cutting-edge LLMs and transformers in DNA cryptography is still a developing field despite these developments. Research and development is still needed to address issues like scalability in real-world implementations, thorough security analysis against LLM- assisted cryptanalysis, and integrating adaptive LLM-driven key generation with biological encoding.

IV. METHODOLOGY

Architecture of the System The suggested system consists of six primary parts and is a modular hybrid framework:

Preprocessor: Transforms tokens into binary sequences after tokenizing plaintext.

DNA Encoder/Mapper: Increases obfuscation by mapping binary sequences to DNA bases (A, C, G, and T) with optional scrambling and shuffling operations.

LLM Key Generator: Takes an initial pseudo-key or contextual input and uses an LLM neural network model to produce final cryptographic keys (in binary or DNA form).

Encryptor: Combines traditional cryptography techniques with cryptographic operations like XOR, DNA-level manipulations, and optional

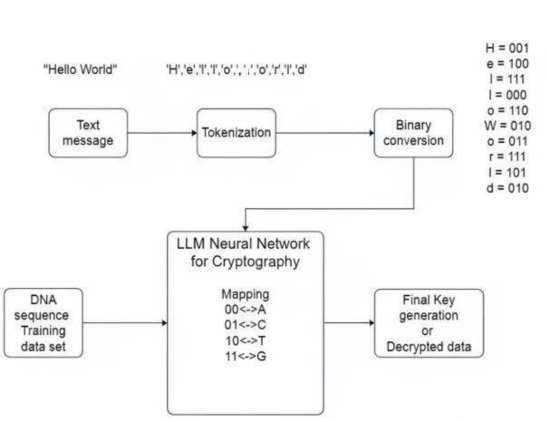


Fig. 1. DNA Cryptography Architecture Proposed Using LLM

Channel/Noise Model: To test robustness, it simulates transmission noise like bit flips, base substitutions, and indels.

Decryptor + LLM Error Corrector: Uses the LLM to generate cryptographic keys, reverses encryption, and performs LLM- assisted decoding and error correction to retrieve the original plaintext.

Procedure for Encryption and Decryption The first step in the encryption process is to transform the input text into one-hot encoded vectors, which are subsequently converted into binary sequences using token indices. A binary-to-DNA encoding scheme (00→A, 01→C, 10→G, 11→T) is used to map binary sequences to corresponding DNA sequences. The final key is obtained by running a randomly generated pseudo-key through the LLM model. In order to create an encrypted DNA sequence for transmission, encryption entails XOR operations between binary DNA representations and the generated key.

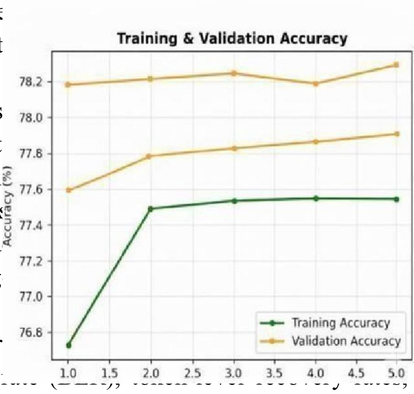
This process is reversed during decryption: the final decryption key is generated by inputting the pseudo-key into the LLM. The original DNA binary data is recovered by XORing the encrypted DNA sequence with the key, decoding it back into one-hot vectors, and then reconstructing it into plaintext tokens.

Different plaintext datasets, such as the 20 Newsgroups corpus for variety, specific Project Gutenberg texts for long-form content, and custom short messages to mimic Internet of Things scenarios, are used in experimental setup experiments. Synthetic DNA sequence ns and publicly accessible nucleotide datasets from NCBI/GenBank support and evaluation for biological realism. Triplets of plaintext, DNA ciphertext, on pipeline make up training datasets, which are enhanced with noisy versions

Systems with Intel i5/i7 processors, at han 256GB are used for computations. Python is used for primary programmin aphy libraries for DNA processing and encryption features. PyTorch and Ten velopment of LLM models. Tools for bioinformatics and visualization includ Aplotlib. With optional access to DNA synthesis/sequencing hardware for long is compatible with Windows, Linux, or macOS.

Baselines and Evaluation Metrics Secur nber of ways:

Correctness and Reliability: Bit error and decryption accuracy (percent full recovery).



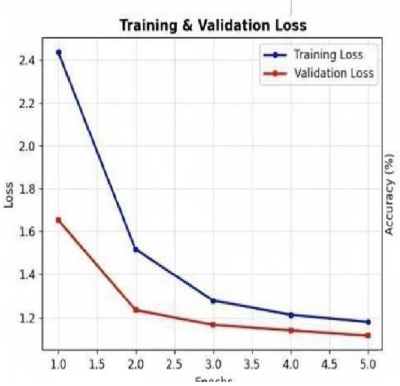
Randomness and Key Quality: Shannon entropy metrics of keys and ciphertexts, NIST Statistical Test Suite facets (frequency, runs, autocorrelation), and avalanche effect measures.

Security and Cryptanalysis: Keyspace entropy, brute-force resistance estimates, statistical attack success rates, and LLM- based cryptanalysis effectiveness as a new threat class.

Performance and Scalability: Throughput in MB/s, end- to-end latency including LLM inference, computational and memory costs, and trade-offs between model size and secu- rity/accuracy gains.

Comparative Baselines: Benchmarked against pure DNA- only schemes, prior LLM-assisted DNA cryptography meth- ods, and hybrid classical+DNA layered encryption frame- works.

V. RESULT ANALYSIS AND DISCUSSION

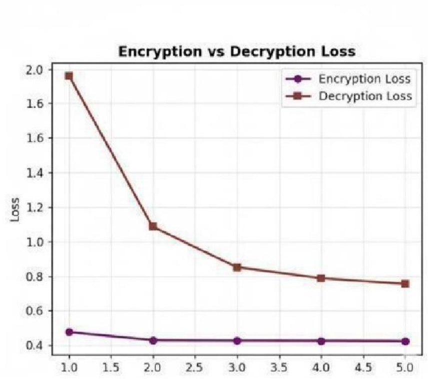


The training and validation loss curves show a consistent downward trend across epochs. Training loss decreased from 2.4 to nearly 1.18, while validation loss reduced from 1.6 to nearly 1.12. This indicates effective learning and convergence of the model. The close gap between training and validation loss suggests that the model does not suffer from overfitting,.

Training accuracy improved from 76.8 to 76.8 to 77.6, while validation accuracy increased from 77.9 to 78.2 percent across epochs. The validation accuracy continuously stayed marginally higher than the training accuracy, demonstrating the LLM's strong handling flexibility and resilience.

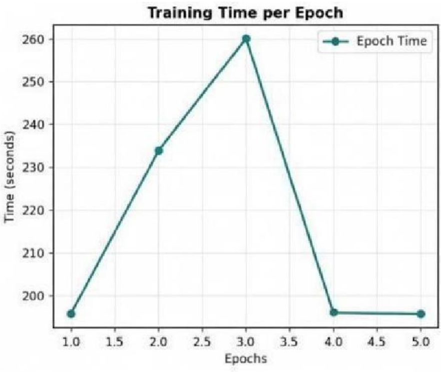
DOI: 10.48175/IJARSCT-30738

Encoded data based on DNA. Additionally, the model appears to have stabilized rapidly based on the marginal improvements across epochs.

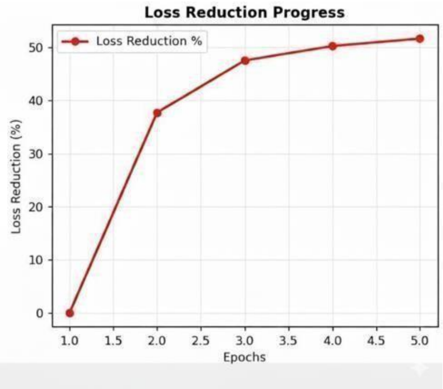


The encryption process confirmed effective learning during the encoding phase by achieving low and stable loss values (0.4 to 0.5). Decryption loss, on the other hand, started out higher at almost 2.0 but gradually decreased to 0.75 by the last epoch.

This discrepancy shows that decryption is still computationally more difficult and introduces small errors. The findings imply that although encryption is very dependable, decryption needs additional optimization to lower residual loss.

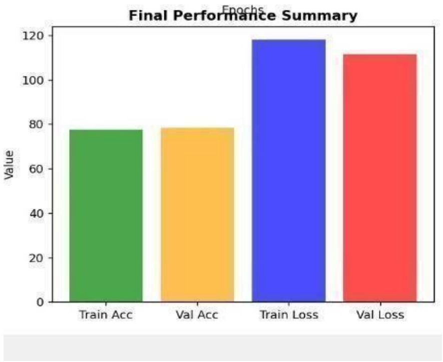


Training time fluctuated greatly, reaching a peak of 260 seconds in Epoch 3 and then stabilizing at 195 seconds in subsequent epochs. Dynamic changes in weight updates and resource distribution during early training are responsible for the initial oscillations. The stabilization shows that the model effectively adjusted over time, lowering computational overhead.



DOI: 10.48175/IJARSCT-30738

Over time, the model's loss decreased, and by the fifth epoch, it had dropped by over half. The majority of the improvement occurred early on, suggesting that the LLM identified the primary DNA-encoding patterns rather rapidly. Smaller, more gradual improvements resulted from the training process's subsequent refinement of what it had already learned.



The bar chart summarizes the overall performance:
Training Accuracy: 77.6
Validation Accuracy: 78.2
Training Loss: 118
Validation Loss: 111
The model's good generalization to new data is further supported by the narrow difference between training and validation metrics. Although predictions are accurate, additional fine-tuning could reduce the loss magnitude, as indicated by the comparatively higher loss values (compared to accuracy).

ACKNOWLEDGMENTS

The Department of Information Science at the Global Academy of Technology, Bangalore, provided the tools and assistance needed to conduct this study, for which the authors are truly grateful. We also thank our faculty mentors for their invaluable advice and insights during the research. Lastly, we would like to thank the open- source research community for their contributions to large language models and DNA cryptography, which provided a solid basis for this work.

REFERENCES

[1]. Digital Data Protection Using Feistel Network and DNA Cryptogra- phy, Aya Goudjil, Lylia Djilali, Amira Benabdelmoumene, M'hamed Hamadouche, Mohamed Amine Riahla, IEEE 2024.

[2]. A Review of DNA Cryptography: From a Data Protection Perspective, Parth Parmar, Jekil Gadhiya, Satvik Vats, Deepak Kumar Verma, Krunal Vaghela, IEEE 2023.

[3]. Image Encryption using DNA Cryptography and Huffman Encoding, Anusha Tripathi, Chittaranjan Pradhan, Archie Bhaumik, IEEE 2024.

[4]. Harnessing DNA Cryptography with the Kyber Algorithm for Enhanced Data Security, Bambang Harjito, Faisal Rahutomo, Dwiko Satriyo. U. Y. S, Heri Prasetyo, IEEE 2023.

[5]. Safeguard Algorithm by Conventional Security with DNA Cryptography Method, Ahmed Hassan Hadi, Sameer Hameed Abdulshaheed, Salim Muhsen Wadi, IEEE 2022.

[6]. Data Encryption and Decryption Using DNA and Embedded Technol- ogy, Manoj Kumar B C, Anil Kumar R J, Shashidhara D, Prem Singh M, IEEE 2022.

[7]. Enhancing Privacy using DNA based Data Hiding, M Shreyansh Narayan, Priyanka Biswas, Nirmalya Kar, IEEE 2024.

[8]. A Hybrid Cryptography Approach Using Symmetric, Asymmetric and DNA Based Encryption, Vikas Yadav, Dr. Manoj Kumar, IEEE 2023.

[9]. A Digital DNA Sequencing Engine for Ransomware Detection Using Machine Learning, Miss.M.Ramya, Avinash Kumar S S, Haresh V, IEEE 2024.

[10]. A Research on DNA and RSA Cryptography for Hybrid Encryption and Decryption for Cloud Processing via IOT Devices, Prashant Bhati, Saurabh Tripathi, Shristi Kumari, Suryansh Sachan, Reena Sharma, IEEE 2023.

[11]. Y. Niu, Z. Zhou, and X. Zhang, "An Image Encryption Approach Based on Chaotic Maps and Genetic Operations," Multimedia Tools and Applications, vol. 79, pp. 26389–26414, 2020, doi: 10.1007/s11042-020-08906-6

[12]. T. Dutta and M. Gupta, "An Intelligent Image Encryption Scheme Based on Hyperchaotic Map and Dynamic DNA Encoding," SN Computer Science, vol. 5, no. 888, Sep. 2024, doi: 10.1007/s42979-024-03224-2.

[13]. Qiuyu Zhang and Jitian Han, "A Novel Coler Image Encryption Algo- rithm Based on Image Hashing, 6D Hyperchaotic and DNA Coding," Multimedia Tools and Applications, vol. 80, no. 9, pp. 13841–13864, Jan. 2021, doi: 10.1007/s11042-020-10437-z.s

DOI: 10.48175/IJARSCT-30738