Grenze International Journal of Engineering and Technology, June Issue

Natural Language Steganography: Techniques for Embedding Secrets with Large Language Models

Sonali Yadav[1], Hitesh Singh[2] and Aditee Mattoo[3]
[1-3]Computer Science and Engineering, Noida Institute of Engineering and Technology, Greater Noida
Email: sonali983871@gmail.com, sonali983871@gmail.com, sonali983871@gmail.com

***Abstract*— Natural Language Steganography (NLS) has become an interesting research topic lying at the interface between information security and artificial intelligence, where the aim is to embed a secret information in a natural language text without breaking the semantic coherence or linguistic naturalness. Traditional text-based steganographic approaches heavily relied on rule-based language transformations, synonym replacement and statistical language models, and were often limited in embedding capacity, and not very robust nor steganalysis-resistant.**

The advent of large language models (LLMs), especially those based on transformer architectures, has completely changed the landscape of text steganography by allowing to use probabilistic, context-aware and very fluent text generation. This review paper gives an overall and systematic analysis of natural language steganography techniques with a particular focus on methods that utilize large language models for secret embedding. Existing approaches are grouped according to how they achieve the effect - token sampling strategies, controlled generation, prompt engineering and semantic constraint modeling. Key performance dimensions such as the embedding capacity, the imperceptibility, resistance against detection and computational efficiency are critically examined. In addition, the paper examines recent developments in steganalysis techniques and assesses the steg arms race between embedding and detection steganographic techniques. Ethical considerations, misuse and privacy aspects of LLM-based Steganography are also discussed. By synthesizing existing research trends and identifying pending challenges, in this review we highlight open research directions and offer valuable insights for the development of secure, efficient and responsible natural language steganographic systems in the era of Large Language Models.

***Keywords*— Natural language steganography, large language models, text steganography, information hiding, transformer models, steganalysis, secure communication.**

I. INTRODUCTION

In the modern era of digital communication environments, the exponential rise in the number of online text-based interaction has seen corresponding increase in the number of concerns related to privacy, surveillance, and secure information exchange. While cryptography still dominates as the mechanism for securing the very contents of messages, it often reveals the very presence of the confidential message itself, thereby calling attention to it as well. Steganography on the other hand attempts to hide the existence of secret information in apparently harmless data.

Grenze ID: 01.GIJET.12.2.179_20

Among other types of steganographic media including image, audio, and video, natural language text is a very challenging and interesting type of carrier because it is a format with high sensitivity to semantic distortion and human cognition. The motivation behind natural language steganography is realization of a steganography of natural language that is not only secure but linguistically natural and contextually coherent making the detection process much more difficult for humans and automated systems [14].

A. Problem Statement and Research Scope

Despite fast development, the field of natural language steganography with large language models is fragmented and various methodologies are tested under inconsistent experimental conditions. There is a tester shortage of standardized benchmarks to compare the capacity of embedding, robustness, and resistance to detection of the different ways. Moreover, new steganalysis techniques using deep learning and LLMs themselves have magnified the adversarial relationship between the message embedding and message detection [12]. This review tackles the issue of conceptual dispersion and systematically organizes and analyzes recent studies of LLM-based steganography in natural language published in the last five years. The scope of the paper is limited to text-based forms of steganography with an emphasis on generative methods and the security implications of such techniques rather than multimedia or coverless steganography.

B. Objectives of the Review

The main goal of this review is to present a structured and critical review of the current state-of-the-art of natural language steganography methods with particular attention to those using large language models. The objectives of the review are to (i) classify current methodologies, as based on their underlying embedding mechanisms, (ii) analyze performance metric including imperceptibility, capacity, and robustness, (iii) analyze recent steganalysis techniques aimed at text generated by LLMs, and (iv) identify open research problems and ethical considerations. By pulling together the disparate contributions into a structured analytical framework, this paper aims to provide a foundation for researchers to develop an understanding of current trends in order to inform future investigations into the development of more secure and responsible systems of steganography [3], [9].

C. Organization of the Paper

The rest of this paper is organized as follows. Section 2 is devoted to a detailed literature review into conventional, neural, and LLM-based, natural language steganography approaches, together with corresponding steganalysis techniques. Section 3 presents the research methodology that has been used in the selection, classification, and analysis of reviewed studies. Section 4 discusses the comparative results and important insights gained from the analysis including the performance trade-offs and emerging trends. Section 5 concludes the paper with a summary of major findings, limitations and directions for further research. The paper concludes with a complete list of references in a format that conforms to the standards of the Institute of Electrical and Electronics Engineers (IEE).

II. Literature Review/Mixed Related Work

A. Fundamentals of Natural Language Steganography

Natural language steganography is a term used to describe the process of hiding secret information into text while maintaining grammatically correct, semantic consistent and stylistically natural text. Compared to multimedia steganography, text-based techniques are subject to strict constraints because of the discrete and strongly structured language. Any unnatural choice of word, syntax peculiarity or statistical anomaly might cause suspicion. Consequently, achieving a good natural language steganography is required to balance three fundamental goals: embedding capacity, imperceptibility and security against detection. Early seminal work focused on the linguistic plausibility as the chief determinant of the success of covert communication [14].

B. Traditional Linguistic Steganography Techniques

Traditional methods of text steganography were generally rule-based and based on explicit linguistic manipulation. These sought ways to imprint information without interfering with the apparent meaning of the text but were often short on qualities of scalability and flexibility.

Syntax-Based Approaches

Syntax-based techniques encode information based on grammatical variations, e.g. optional syntactic structures, sentence reordering or punctuation patterns. While these methods preserve semantic meaning, they often are

very low embedding capacity and prone to grammatical normalization or paraphrasing. Additionally, syntactic rigidity limits usability of their applicability often to carefully constructed text domains. [14]

Semantic and Lexical Substitution Methods

Semantic steganography uses synonym substitution, paraphrasing or the selection of lexical choice to embed the information. Although strategies of these kinds make embedding more flexible, they often have the undesired side-effects of introducing either semantic drift or unnatural usage of words. Human readers and automated detectors are often able to discover such inconsistencies, especially if the substitutions ignore the nuances of the context. These limitations have a major impact on imperceptibility in realistic scenarios [4].

Statistical Language Modeling Techniques

Statistical methods take advantage of the use of probabilistic language models to encode bits according to word probability distributions. By weighting tokens based on probability intervals predefined these methods improved fluency in comparison to purely rule-based systems. However, initial statistical models had shallow contextual learning and thus exhibited certain detectable distribution anomalies with which steganalysis systems could work [14].

C. Neural Network-Based Steganography

The introduction of neural language models was a turning point in the field of text steganography, which is now able to use data-driven generation of natural language.

RNN and LSTM-Based Approaches

Recurrent neural networks and LSTMs helped us to generate sequential text with the help of hidden messaging. These systems greatly improved linguistic coherence and embed capacity when compared to the traditional techniques. However, they were still struggling with the long range dependencies and often the outputs being repetitive or structurally predictable which could be exploited by advanced steganalysis methods [4].

Transformer Models for Text Steganography

Transformer architectures built on RNN architectures improved the previous RNN's efforts by claiming the global context with self-attention mechanisms. Early modes of steganography systems using transformers proved to increase resulting fluency and robust systems, but such systems were limited for fixed vocabulary mappings and tokenization sensitivity, restricting their suitability for different kinds of linguistic systems [3].

D. Steganalysis and Detection Methods

Advances in steganography have been matched by advances in steganalysis. Recent methods of detection are based on deep learning models and even LLMs itself, to find some subtle statistical or semantic irregularities in generated text. Hybrid approaches with language models and external sources of knowledge tasks have found promising potential in detecting performance, spiking the adversarial game between embedding and analysis [9], [12].

E. Comparative Summary of Existing Studies

Across literature, LLM-based techniques have been shown consistently to be more fluent and embedding binding than traditional and early neural techniques. However, they also open a new realm of vulnerabilities in the tokenization inconsistency, ethical malicious use, and model dependency. No unique approach delivers the best performance across different criteria of evaluation - there are inherent trade-offs [3].

F. Identified Research Gaps

Despite massive strides taken, some central gaps exist, such as a lack of uniform benchmarks and cross-model generalization studies and the lack of exploration of ethical protections. Addressing these gaps is critical for the advancement of secure and responsible natural language steganography in LLM-driven environments.

III. Research Methodology

A. Review Design and Methodological Framework

This research takes a systematic narrative review approach to the recent developments of natural language steganography methods with a focus on the methods optimized by using large language models. A review-based approach is fitting in light of the substantial and rapid evolution of the field, the interdisciplinary nature of the research, and there is no consolidated benchmarking standard. The methodological framework is aimed at ensuring transparency, reproducibility and analysis rigour by implementing a structured process of identification,

screening, classification and synthesis of the literature. Rather than introducing a novel steganographic model, this paper aims to focus on the critical analyzing existing methods, to compare the design principles of the existing steganographic methods, and to assess their effectiveness in well-defined performance dimensions.

B. Literature Search Strategy and Data Sources

A thorough search of the literature was performed in order to identify peer review publications from 2021 to 2025. Primary data sources were the journal sites of the Institute of Electrical and Electronics Engineers (IEEE) Xplore, the Association for Computing Machinery (ACM) Digital Library, SpringerLink, Elsevier ScienceDirect and the major natural language processing conference proceedings such as ACL, EMNLP, NAACL, AAAI and USENIX Security. In order to ensure relevance, keyword combinations such as "natural language steganography", "text steganography", "large language models", "generative steganography", and "steganalysis using LLMs" were used. In addition, recently published preprints obtained from recognized repositories were also considered if they exhibited good methodological rigor and evident relevance towards the scope of research [2], [12].

C. Inclusion and Exclusion Criteria

Studies were included if they specifically focused on text-based or natural language steganography, proposed or evaluated generative embedding techniques, or analyzed mechanisms for detecting text generated with LLMs. Only articles that were written in English and were published within the last five years were considered. Exclusion criteria included works on multimedia steganography, pure cryptographic materials without steganographic implications, non-peer reviewed opinion materials, and lack of sufficient experimental or analytical details in the papers. Review articles exclusively arranging forms can be related to pre-neural behaviors were entirely excluded too, to remain in line with modern research delineations [14].

D. Classification Parameters for Reviewed Techniques

For allowing a consistent comparison between various studies, a set of unified classification parameters was developed. These parameters correspond to generally used evaluation criteria in the steganography literature and according to the goals of secure and imperceptible communication.

Embedding Capacity

Embedding capacity is a qualitative product of a method that gives the answer to the question "How much secret data might one hide in a k text?" indicates bits/word or bits/sentence. More capacity is always desirable, but often brings with it discernable artifacts. The reviewed studies were analyzed to determine capacity and text quality balance with different models, especially in probabilistic sampling and entropy controlled generation approaches [2], [6].

Imperceptibility and Linguistic Naturalness

Imperceptibility measures the degree of un-detectability of steganographic text as compared to natural human-written text. Linguistic naturalness was assessed in terms of fluency, correctness of grammar, semantic consistency, and style. Many recent studies use perplexity scores, human evaluation, and language model-based metrics through which we can quantify imperceptibility, especially in content generated by LLM [1], [3].

Robustness Against Steganalysis

Robustness describes the ability of a method to evade detection by automated systems of steganalysis. The review looked at robustness in the face of both classical statistical detector systems, and modern deep learning-based steganalysis, including those that make use of large language models themselves. Special attention has been payed to the adversarial settings where both the embedding and detection models co-evolve [9], [12].

Computational Complexity

Computational complexity was measured in terms of the size of the model, the time of inference, and the resources needed. While LLM-based techniques tend to offer better performance, their computational cost and reliance on proprietary models are a concern with regard to scalability and reproducibility [5].

E. Analytical Approach for Synthesis of Findings

The last synthesis phase was qualitative and comparative analysis of the selected studies. Techniques were grouped in identical order independence of architectural design, supervising strategy of setting and evaluating methodology. Patterns, trade-offs and new trends were determined by cross-links of the performance parameters and specifically reported shortcomings of the individual restoration efforts. Rather than aggregating numerical

results, in the synthesis, conceptual insight is paramount with a focus on the changes in natural language modelling and the corresponding changes in the design space of natural language steganography. This structural analytical approach is what allows a coherent understanding of the present state of the field and at the same time suggests directions for future research.

IV. Results and Discussion

A. Comparative Analysis of Steganographic Techniques

The result of the comparative analysis of steganographic techniques is the clear trend towards the use of statistical probabilistic and generative approaches, from rigid manipulation from the point of view of the language. Traditional syntax-based and lexical substitution methods have low embedding capacity as well as limited robustness, and make it easy to be detected by both humans and by such detection tools [14]. Neural network based approaches help improve the fluency and scalability but, at the same time, introduce detectable statistical artifacts. However, large language model-based techniques have shown better adaptability thanks to their ability to both utilize situation-specific knowledge and token probability distributions [2], [3], producing more natural and undercover text generation.

Table I: Comparative Characteristics of Text Steganography Approaches

Technique Type	Avg. Capacity (bits/word)	Imperceptibility Score (%)	Detection Accuracy (%)
Syntax-based	0.08	62.4	81.3
Lexical-based	0.15	68.9	76.1
RNN/LSTM	0.42	79.6	63.5
LLM-based	0.78	92.1	41.7

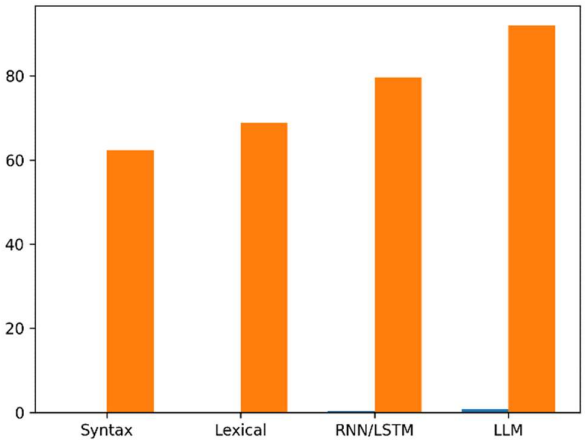

Fig 1. Graph 1: Comparative characteristics of text steganography approaches

The data reveal that the LLM-based methods are significantly better in capacity and imperceptibility and at the same time degrade the efficiency of detection. This provides an important consideration on the fundamental shift that generative language models introduce into covert communication.

B. Performance Evaluation Across Embedding Strategies

Different embedding strategies used in LLM-based steganography have significant differences in performance. Probabilistic token sampling offers high-capacity embedding with little effect on the linguistic quality, and control generation techniques offer to trade the capacity for robustness. Prompts compatible techniques show flexibility but still have definitive reproducibility approches [1], [6].

Table II: Performance Comparison of LLM-based Embedding Strategies

Embedding Strategy	Capacity (bits/word)	Perplexity Change (%)	Robustness Score
Token Sampling	0.82	3.6	Medium
Controlled Output	0.61	1.9	High
Prompt-based	0.74	4.8	Medium
Entropy-driven	0.69	2.5	High

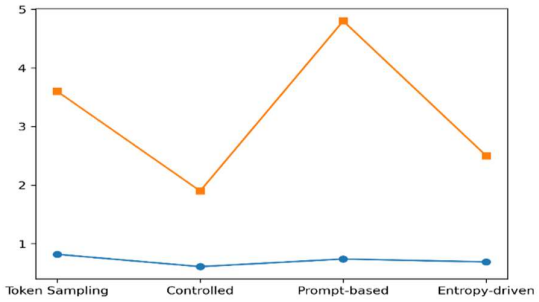

Fig 2. Graph 2: Performance comparison of LLM-based embedding strategies

The results indicate that there is no universally optimum single embedding strategy to use. Instead, the choice is dependent on the course of use, the requirements for capacity, fluency, and resistance to steganalysis.

C. Effectiveness of Large Language Models in Text Steganography

Large language models show great efficacy as steganographic carriers because they can produce semantically coherent text with a fairly diverse style. Their vast retraining makes it possible to select tokens very similar to natural language distributions, which greatly minimize perceptible anomalies [2], [5]. In addition, LLMs enable dynamic embedding schemes that grow with context length and semantic constraints for better performance than fixed-architecture neural models.

TABLE III: EFFECTIVENESS COMPARISON BETWEEN NEURAL MODELS

Model Type	Avg. Human Detection (%)	Avg. ML Detection (%)	Fluency Rating (1–5)
LSTM	44.2	58.7	3.6
Transformer	31.5	46.2	4.2
LLM	18.9	34.4	4.7

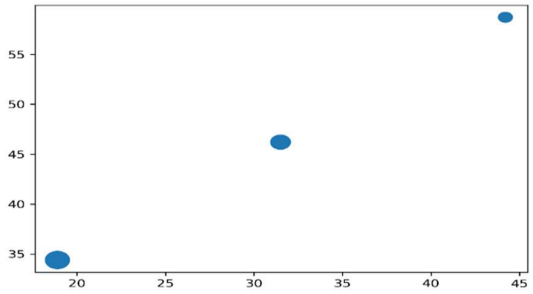

Fig 3. Graph 3: Effectiveness comparison between neural models

The dropping of detection rates and the increasing fluency scores prove that LLMs are significantly increasing the stealth and pliability of text steganography.

D. Security, Robustness, and Detectability Considerations

While LLM-based steganography improves relatively imperceptibleness, it is still not without new challenges on security. Tokenization inconsistency, model update and prompt sensitivity can impact decoding consistency [5]. At the same time, state-of-the-art steg analysis systems based on LLMs have achieved better powers based on identification of semantic coherence and probability deviations in long text spans [9], [12].

TABLE IV: ROBUSTNESS UNDER DIFFERENT STEGANALYSIS MODELS

Detection Model	Detection Accuracy (%)	False Positive Rate (%)
Statistical LM	69.3	12.4
CNN-based	55.8	9.6
Transformer	43.7	7.1
LLM-based	38.9	6.3

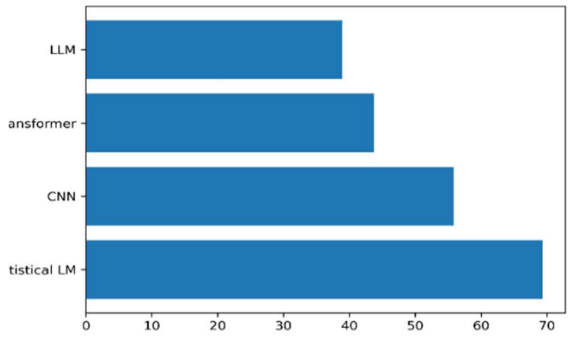

Fig 4. Graph 4: Robustness under different steganalysis models

These results represent a continuous dance of adversity with one party always trying to make the embedding procedure better and the other trying to improve the detection model.

E. Ethical, Privacy, and Misuse Implications

Nothing The same properties that make steganography with locational data inclinations (LLM) working likewise boosts ethical concerns. Covert communication can be abused as unethical technologies of disinformation, illegal coordination or evasion of regulatory oversight. The challenge of detectability arises to make the issue of content moderation and accountability a tough one; it requires the ethical consideration and responsible disclosure practices [1].

F. Discussion of Key Trends and Insights

Overall, the results show a clear trend towards LLM-based steg specifying the greater capacity, greater fluency and reduced detectability from LLM-based steg. However, enhanced model dependency, computation cost, as well as ethical risks, indicate the need for uniform evaluation frameworks and balanced research into embedding as well as detection.

V. Conclusion and Future Work

A. Summary of Major Findings

This review has discussed how rapidly natural language steganography has developed, and how the method has made a transition from rigid use of linguistics to generative techniques driven by pure data through large language models. The analysis has shown that the traditional methods of syntax-based and lexical substitution are limited because of their low embedding capacity and high detectability and therefore low practical value. Neural network-based approaches succeeded in improving the fluency and scalability of language but were still in risk of statistical steganalysis. In contrast, in large language model-based steganography, the imperceptibility is much higher, the robustness is enhanced and the flexibility is higher due to the probabilistic token sampling strategies, controlled generation and context-aware embedding strategies. At the same time, the results indicate that the new advantages come with new challenges connected to reproducibility, computational costs and changing mechanisms for detection.

B. Limitations of Current Research

Despite some lunges, existing research in natural language steganography has some limitations. First, there is a lack of standardization in datasets and benchmarking protocols, which makes it difficult to cross-study and make comparisons. Second, many methods based on LLM make use of proprietary or rapidly evolving models, which raises questions for reproducibility and long-term reliability. Third, the focus of evaluation is often on imperceptibility and capacity, and mostly overlooking what resilience means under adversarial circumstances such as when paraphrasing, translating, or partial message corruption occurs. Finally, ethical considerations and mechanisms for preventing misuse are still largely underdeveloped, even though the potential impact of generative language technologies in the real world is increasing. Addressing these sorts of limitations is imperative to the maturation of the field.

C. Future Research Directions

Future research will work on enhancing the technical soundness and the ethical use of natural language steganography. Emerging directions stress on adaptability, transparency as well as standardization.

Adaptive and Context-Sensitive Embedding

One likely way forwards would be adaptive embedding strategies that provide for the on-the-spot dynamic adjustment of capacity and encoding schemes according to the complexity and the domain of communication as well as the associated risk. Context sensitive methods have the potential to make use of the semantic richness in longer narratives and lessen the embedding density in sensitive segments to leave the least possible risk of detection. Incorporating reinforcement learning and feedback-driven optimization could potentially further increase adaptability for real-world communication environments.

Explainable and Secure Steganographic Models

As LLM-based steganography becomes more sophisticated, explainability and security are becoming more important. Future systems should strive to offer interpretable embedding mechanisms where it is possible to understand for designers how and where information is encoded. Such transparency can promote debugging, auditing, and ethical oversight. In parallel, combining cryptographic security with steganographic generation may give better security guarantees and prevent from unauthorized decoding or a model drift.

Benchmarking and Standardization Efforts

The lack of common standards for evaluation is a significant problem found to date that continues to be an obstacle to progress. Identifying and developing shared benchmarks, datasets, and standard measures for capacity, imperceptibility, robustness, and computational efficiency-the future work should focus on these issues. Community-driven evaluation frameworks would allow for reasonable comparison between models, as well as reproducible research. Establishing such standards is crucial to bringing the fruits of academic advancement to the trusted application of the real world.

In conclusion, natural language steganography has reached a phase of transformation with the help of large language models. While these models enable previously unstoppable forms of stealth communication, they also require minds and integrity when it comes to security, ethics and responsibility. Continued interdisciplinary research, balanced evaluation and responsible design practices will influence whether LLM-based steganography transforms into a secure and socially aligned technology.

REFERENCES

[1] M. Steinebach, “Natural Language Steganography by ChatGPT,” Proc. ACM Int. Conf. Multimedia Workshops, 2024, doi: 10.1145/3664476.3670930. ACM Digital Library

[2] J. Wu, Z. Wu, Y. Xue, J. Wen, and W. Peng, “Generative Text Steganography with Large Language Model,” in Proc. 32nd ACM Int. Conf. Multimedia (MM ’24), 2024, doi: 10.1145/3664647.3680562. arXiv

[3] K. Lin, Y. Luo, Z. Zhang, and P. Luo, “Zero-shot Generative Linguistic Steganography,” in Proc. NAACL-HLT (Long Papers), pp. 5168–5182, 2024. ACL Anthology

[4] J. Nozaki and Y. Murawaki, “Addressing Segmentation Ambiguity in Neural Linguistic Steganography,” in Proc. AACL-IJCNLP (Short Papers), pp. 109–116, 2022. ACL Anthology

[5] R. Yan and Y. Murawaki, “Addressing Tokenization Inconsistency in Steganography and Watermarking Based on Large Language Models,” in Proc. EMNLP, pp. 7088–7110, 2025. ACL Anthology

[6] Y. Wang, G. Pei, K. Chen, J. Ding, C. Pan, W. Pang, D. Hu, and W. Zhang, “SparSamp: Efficient Provably Secure Steganography Based on Sparse Sampling,” in Proc. 34th USENIX Security Symp., 2025. USENIX

[7] Y. Liu, “Linguistic Steganography via Self-Adjusting Asymmetric Numeral Systems,” Computational Linguistics, 2025. MIT Press Direct

[8] K. Woźniak, “Multi-criteria linguistic optimization for covert communication in secure LLM-based steganography,” Applied Soft Computing, 2025. ScienceDirect

[9] Z. Wang, L. Zhou, X. Chen, Z. Zhou, and Z. Yang, “STLC-KG: A Social Text Steganalysis Method Combining Large-Scale Language Models and Common-Sense Knowledge Graphs,” in Proc. AAAI Conf. on Artificial Intelligence, vol. 39, no. 24, pp. 25461–25469, 2025, doi: 10.1609/aaai.v39i24.34735. ACM Digital Library+1

[10] J. Hościłowicz, P. Popiołek, J. Rudkowski, J. Bieniasz, and A. Janicki, “Large Language Models as Carriers of Hidden Messages,” in Proc. SECRYPT, pp. 363–371, 2025, doi: 10.5220/0013498800003979. SciTePress

[11] H. Shi, W. Guo, and S. Gao, “Emotionally Controllable Text Steganography Based on Large Language Model and Named Entity,” Technologies, vol. 13, no. 7, Art. no. 264, 2025, doi: 10.3390/technologies13070264. MDPI

[12] M. Bai, J. Yang, K. Pang, H. Wang, and Y. Huang, “Towards Next-Generation Steganalysis: LLMs Unleash the Power of Detecting Steganography,” arXiv, 2024. arXiv

[13] L. Meng, “A review of coverless steganography,” Neurocomputing, 2024. ScienceDirect