

Enhancing Contextual Compatibility of Textual Steganography Systems Based on Large Language Models

Nasouh AlOlabi

Higher Institute for Applied Sciences and Technology
Damascus, Syria

Riad Sonbol

Higher Institute for Applied Sciences and Technology
Damascus, Syria

ABSTRACT

This systematic literature review examines the transformative impact of Large Language Models (LLMs) on linguistic steganography. Through comprehensive analysis of current research, we demonstrate that LLM-based approaches significantly enhance imperceptibility, embedding capacity, and naturalness in cover text generation, addressing traditional limitations of low embedding capacity and cognitive imperceptibility. Our findings reveal a paradigm shift towards context-aware steganographic systems that leverage domain-specific knowledge and communicative context to achieve both perceptual and statistical imperceptibility. The review establishes that understanding contextual compatibility and domain correlations is crucial for developing more sophisticated, robust, and secure covert communication systems, paving the way for future advances in generative text steganography.

KEYWORDS

Systematic Literature Review, Linguistic Steganography, Large Language Models, LLMs, Natural Language Processing, NLP, Software Engineering

1 INTRODUCTION

This review explores how large language models (LLMs) are transforming linguistic steganography, the practice of hiding messages in text. We focus on the unique challenges and advances in using LLMs for secure, imperceptible, and high-capacity covert communication.

1.1 Overview of Information Security and Concealment Systems

Information security systems include **encryption**, **privacy**, and **concealment** (steganography).

1.1.1 Encryption Systems and Privacy Systems. These protect content but reveal that secret communication is happening, which can attract attention.

1.1.2 Concealment Systems (Steganography). Steganography hides the existence of information by embedding it in ordinary carriers (e.g., text, images). The fundamental goal is to achieve **imperceptibility**, which encompasses three dimensions: **perceptual imperceptibility** (the steganographic text appears natural and indistinguishable from normal text to human observers), **statistical imperceptibility** (the statistical properties of the steganographic text match those of the cover medium), and **cognitive imperceptibility** (the semantic content and contextual coherence remain consistent with expected communication patterns). Text is a challenging carrier due to its low redundancy and strict semantics.

1.2 Introduction to Steganography

Steganography is often explained by the “Prisoners’ Problem,” where Alice and Bob must communicate secretly under surveillance. The goal is to embed messages so they are undetectable to an observer.

Steganography methods include **carrier selection**, **carrier modification**, and **carrier generation**.

- **Carrier modification:** Hide information in existing text with minimal changes.
- **Carrier generation:** Generate new text that encodes information, allowing higher capacity but requiring naturalness.

1.3 The Significance of Linguistic Steganography

Linguistic steganography enables covert communication, especially where encryption is suspicious. Text is a robust, ubiquitous carrier but presents challenges in balancing imperceptibility and capacity. Advances in deep learning and LLMs improve text quality and security, while related fields like watermarking focus on tracing content origin.

1.4 Key Terminology and Definitions

To ensure accessibility for readers from diverse academic backgrounds, we provide formal definitions of critical technical terms used throughout this review:

- **Perceptual Imperceptibility:** The property that steganographic text appears natural and indistinguishable from normal text to human observers, maintaining linguistic fluency and contextual appropriateness.
- **Statistical Imperceptibility:** The property that the statistical characteristics of steganographic text match those of the cover medium, making it undetectable by automated statistical analysis.
- **Cognitive Imperceptibility:** The property that the semantic content and contextual coherence of steganographic text remain consistent with expected communication patterns and domain-specific knowledge.
- **Channel Entropy:** A measure of uncertainty or randomness in the communication medium that determines the theoretical capacity for information hiding. Higher entropy allows for greater embedding capacity.
- **Perfect Samplers:** Algorithms that can generate samples from a probability distribution with perfect accuracy, ensuring no statistical deviation from the target distribution—a requirement for provably secure steganography.
- **Explicit Data Distributions:** Clearly defined mathematical representations of the probability distributions governing

the cover medium, enabling precise security analysis and theoretical guarantees.

- **Hallucinations (in LLMs):** Instances where language models generate plausible-sounding but factually incorrect, nonsensical, or contextually inappropriate content due to limitations in training data or model architecture.
- **Psic Effect:** The Perceptual-Statistical Imperceptibility Conflict Effect, representing the fundamental trade-off where optimizations for perceptual quality may compromise statistical security and vice versa.

1.5 Scope of the Review

This review covers LLM-based linguistic steganography, focusing on methods, evaluation, challenges, and future directions.

2 STEGANOGRAPHY AND LARGE LANGUAGE MODELS

Large Language Models (LLMs) have emerged as a significant development in the field of natural language processing, profoundly impacting text generation and related applications like steganography and watermarking. Here's a breakdown of their emergence and impact:

2.1 Capabilities and Approximating Natural Communication

LLMs are **generative models** that can **approximate complex distributions like text-based communication**. They represent the best-known technique for this task. These models operate by taking context and parameters to output an explicit probability distribution over the next token (e.g., a character or a word). The next token is typically sampled randomly from this distribution, and the process repeats to generate output of a desired length. Training LLMs involves processing vast amounts of data to set parameters and structure, enabling their output distributions to approximate true distributions in the training data. The **quality of content generated by generative models is impressive** and continues to improve. This has led to LLMs blurring the boundary of high-quality text generation between humans and machines. LLMs are increasingly used to generate data for specific tasks, such as tabular data, relational triples, sentence pairs, and instruction data, often achieving satisfactory generation quality in zero-shot learning for specific subject categories. They have also shown capabilities in mimicking language styles and semantics, and their generalization ability allows them to comprehend the semantics of context.

2.2 Role in Generative Linguistic Steganography

LLMs are considered **favorable for generative text steganography** due to their ability to generate high-quality text. Researchers propose using generative models as steganographic samplers to embed messages into realistic communication distributions, such as text. This approach marks a departure from prior steganographic work, motivated by the public availability of high-quality models and significant efficiency gains. LLMs like **GPT-2**, **LLaMA**, and **Baichuan2** are commonly used as basic generative models for steganography. Existing methods often utilize a language model

and steganographic mapping, where secret messages are embedded by establishing a mapping between binary bits and the sampling probability of words within the training vocabulary. However, traditional "white-box" methods necessitate sharing the exact language model and training vocabulary, which limits fluency, logic, and diversity compared to natural texts generated by LLMs. These methods also inevitably alter the sampling probability distribution, thereby posing security risks. New approaches, such as **LLM-Stega**, explore **black-box generative text steganography using the user interfaces (UIs) of LLMs**, thereby circumventing the requirement to access internal sampling distributions. This method constructs a keyword set and employs an encrypted steganographic mapping for embedding, proposing an optimization mechanism based on reject sampling for accurate extraction and rich semantics [14]. Another framework, **Co-Stega**, leverages LLMs to address the challenge of low capacity in social media by expanding the text space for hiding messages (through context retrieval) and **increasing the generated text's entropy via specific prompts** to enhance embedding capacity. This approach also aims to maintain text quality, fluency, and relevance [7]. The concept of **zero-shot linguistic steganography** with LLMs utilizes in-context learning, where samples of covert text are used as context to generate more intelligible stegotext using a question-answer (QA) paradigm [8]. LLMs are also employed in approaches like **ALiSa**, which directly conceals token-level secret messages in seemingly natural steganographic text generated by off-the-shelf BERT models equipped with Gibbs sampling [16]. The increasing popularity of deep generative models has made it feasible for provably secure steganography to be applied in real-world scenarios, as they fulfill requirements for **perfect samplers** (algorithms that can generate samples from a probability distribution with perfect accuracy, ensuring no statistical deviation from the target distribution) and **explicit data distributions** (clearly defined mathematical representations of the probability distributions governing the cover medium, enabling precise security analysis) [3, 5, 10].

2.3 LLM-Based Steganography Models

2.3.1 Evaluation Metrics.

Imperceptibility Metrics. Perceptual metrics include PPL, Distinct-n, MAUVE, and human evaluation. Statistical metrics include KLD, JSD, anti-steganalysis accuracy, and semantic similarity.

Embedding Capacity Metrics. Metrics include bits per token/word and embedding rate.

2.4 Challenges and Limitations in Steganography with LLMs

2.4.1 Perceptual vs. Statistical Imperceptibility (Psic Effect). The **Psic Effect** (Perceptual-Statistical Imperceptibility Conflict Effect) represents a fundamental trade-off in steganographic systems where improving perceptual quality can reduce statistical security, and vice versa. This occurs because optimizations that make text appear more natural to human observers (perceptual imperceptibility) may inadvertently introduce statistical anomalies detectable by automated analysis, while techniques that preserve statistical properties

may produce text that appears unnatural or suspicious to human readers.

2.4.2 Low Embedding Capacity. Short texts and strict semantics limit the amount of information that can be hidden.

2.4.3 Lack of Semantic Control and Contextual Consistency. Ensuring generated text matches intended meaning and context is difficult.

2.4.4 Challenges with LLMs in Steganography. LLMs may introduce unpredictability, bias, or leak information.

2.4.5 Segmentation Ambiguity. Tokenization can cause ambiguity in how information is embedded or extracted.

A primary challenge in steganography, particularly when utilizing Large Language Models (LLMs), revolves around the **distinction between white-box and black-box access**. Most current advanced generative text steganographic methods operate under a "white-box" paradigm, meaning they require direct access to the LLM's internal components, such as its training vocabulary and the sampling probabilities of words. This presents a significant limitation because many state-of-the-art LLMs are proprietary and are accessed by users primarily through black-box APIs or user interfaces [14]. Consequently, these white-box methods are often impractical for real-world deployment with popular commercial LLMs. Furthermore, methods that rely on modifying the sampling probability distribution to embed secret messages inherently introduce security risks because they alter the original distribution, making the steganographic text statistically distinguishable from normal text [3, 5, 14, 15].

Another significant hurdle is **ensuring both the quality and imperceptibility of the generated text**, encompassing perceptual, statistical, and cognitive imperceptibility. While advancements in deep neural networks have improved text fluency and embedding capacity, older models or certain embedding strategies can still produce texts that lack naturalness, logical coherence, or diversity compared to human-written content. Linguistic steganography methods often struggle to control the semantics and contextual characteristics of the generated text, leading to a decline in its "cognitive-imperceptibility" [1, 15]. This can make concealed messages easier for human or machine supervisors to detect. Although models like NMT-Stega and Hi-Stega aim to maintain semantic and contextual consistency by leveraging source texts or social media contexts, this remains a complex challenge [1, 13].

Channel entropy requirements and variability also pose a considerable challenge. **Channel entropy** refers to the measure of uncertainty or randomness in the communication medium, which determines the theoretical capacity for information hiding. **Traditional universal steganographic schemes** (general-purpose steganographic methods designed to work across different types of cover media without requiring medium-specific adaptations) often demand that the communication channel maintains a minimum level of entropy, which is rarely consistent in real-world communication, especially in natural language. Moments of low or zero entropy can cause existing steganographic protocols to fail or necessitate the generation of extraordinarily long steganographic texts, making covert communication impractical. While schemes like Meteor attempt to adapt by fluidly changing the encoding rate

proportional to instantaneous entropy, overcoming this variability without increasing detectability is difficult. The "Psic Effect" (Perceptual-Statistical Imperceptibility Conflict Effect) highlights this dilemma, where optimizing for perceived quality might compromise statistical imperceptibility and vice-versa.

Furthermore, **segmentation ambiguity** introduced by subword-based language models, commonly used in high-performing Transformer architectures, presents a critical issue for provably secure linguistic steganography. When a sender detokenizes generated subword sequences into a continuous text (e.g., "any" + "thing" becoming "anything") before transmission, the receiver might re-tokenize it differently (e.g., as a single "anything" token), leading to decoding errors and affecting subsequent probability distributions. Existing disambiguation solutions typically involve modifying the token candidate pool or probability distributions, which renders them incompatible with the strict requirements of provably secure steganography that demand unchanged distributions [10]. While SyncPool attempts to address this without altering the distribution, it may still lead to a reduction in the embedding rate due to information loss [10].

Additional limitations include the following: * **Computational Overhead:** LLMs, while powerful, incur a higher computational cost (3-5 times more than prior methods), which could impact real-time communication scenarios [8]. * **Data Integrity and Reversibility:** Some linguistic steganography methods are not reversible, meaning the original cover text cannot be perfectly recovered after message extraction, which is undesirable for sensitive applications [11, 18]. Text data is generally less prone to lossy compression issues than other media, but incompleteness of the steganographic text can still damage the embedded bitstream [8]. * **Ethical Concerns:** The use of pre-trained LLMs may inadvertently introduce ethical issues such as political biases, gender discrimination, or the generation of insulting content [8]. * **Provable Security and Rigor:** Despite decades of research into **provably secure steganography** (steganographic systems with formal mathematical guarantees that they are indistinguishable from the cover medium under specified assumptions), practical systems have been hampered by strict requirements like perfect samplers and explicit data distributions [3, 5]. Many works from the NLP community, while generating convincing text, often lack rigorous security analyses and fail to meet formal cryptographic definitions, making them vulnerable to detection [5].

Despite their capabilities, generative models are still **far from perfect** in imitating real communication. A significant challenge for practical steganography is the difficulty of finding samplers for non-trivial distributions like the English language, which continues to evolve. When using approximate samplers, there is a risk that an adversary can detect a steganographic message by distinguishing between the real channel and the approximation [5]. LLMs are known to make mistakes, including **hallucinations** (instances where the model generates plausible-sounding but factually incorrect, nonsensical, or contextually inappropriate content due to limitations in training data or model architecture), which can lead to errors and erratic embedding during text generation, especially for long stego sequences. One critical issue is **segmentation ambiguity** in neural linguistic steganography. LLMs often use **subword tokenization**, meaning a single text can correspond to multiple

token representations. If the sender and receiver have different understandings of segmentation, it can lead to incorrect message extraction and affect subsequent generation steps. Current provably secure methods have largely overlooked this. SyncPool is a proposed method to address this by grouping tokens with prefix relationships in the candidate pool without altering the original probability distribution. The **computational overhead of LLMs is higher** compared to prior methods (approximately 3x to 5x), potentially limiting real-time communication. The effectiveness of LLM-based steganography can be limited by the **entropy of the cover text** in social media contexts, as short, context-dependent replies have lower entropy, thus limiting hiding capacity [7].

3 LITERATURE REVIEW METHODOLOGY

3.1 Research questions

Here are the research questions addressed in this SLR:

- What is the state of published literature on steganographic techniques that leverage large language models (LLMs)?
- In which applications are steganographic techniques with LLMs being explored?
- What metrics and evaluation methods are used to assess the performance of steganographic techniques in LLMs, focusing on factors like capacity, security, and contextual compatibility?
- How are external knowledge sources (semantic resources) integrated into steganographic techniques with LLMs to enhance capacity or contextual relevance?
- What are the limitations and trade-offs associated with current steganographic techniques using LLMs, particularly concerning security, capacity, and contextual compatibility?
- What are the potential future research directions in steganography with LLMs, considering emerging trends and identified gaps in the literature?

3.2 Search query string

We used the following search query string for our initial literature search:

(steganography or watermark or "Information Hiding") and ("Large Language Model" or LLM or BERT or LAMA or GPT)

3.3 Study selection and quality assessment

We established the following inclusion and exclusion criteria for study selection:

3.3.1 Inclusion Criteria.

- **Full Text Access:** Studies for which the full text is available.
- **Language:** Publications written in English.
- **Peer-reviewed:** Articles published in peer-reviewed journals, conferences, or workshops.
- **Publication Date:** Studies published from 2018 onwards, to focus on recent advancements in LLMs.
- **Relevance:** Studies directly addressing steganography, watermarking, or information hiding techniques that utilize or are significantly impacted by Large Language Models (LLMs), BERT, LAMA, or GPT architectures.

- **Research Type:** Empirical studies, surveys, reviews, and theoretical contributions.

3.3.2 Exclusion Criteria.

- **Duplicated Studies:** Multiple publications reporting the same study will be excluded, with the most complete or recent version retained.
- **Incomplete or Abstract-only:** Studies for which only an abstract is available or the full text is incomplete.
- **Irrelevant Studies:** Publications not directly related to steganography with LLMs.
- **Non-English Publications:** Studies not published in English.
- **Non-peer-reviewed Sources:** Preprints, dissertations, theses, books, and book chapters (unless they are extended versions of peer-reviewed conference papers).

3.4 Bibliometric analysis

Briefly note if snowballing was used for additional sources.

4 CONDUCTING THE SEARCH

This section details the systematic process followed to identify and select relevant literature for this review. The search strategy was designed to ensure comprehensive coverage of the topic while adhering to predefined inclusion and exclusion criteria.

4.1 Initial Candidate Papers

Our initial automated search across selected digital libraries yielded a total of 1043 candidate papers. The distribution of these papers by source was as follows: ACM Digital Library (346), IEEE Digital Library (61), Science@Direct (209), Scopus (151), and Springer Link (276). This stage focused on broad keyword matching to capture all potentially relevant studies.

4.2 Duplicate Removal

Following the initial search, a rigorous process of duplicate removal was undertaken. After removing duplicates, 989 papers remained. This involved both automated tools and manual verification to ensure that each unique paper was considered only once, thereby streamlining the subsequent screening stages.

4.3 Multi-stage Filtering

The identified papers underwent a multi-stage filtering process based on their titles, abstracts, and full texts. After title and abstract filtering, 58 papers remained. Of these, 18 were accepted with PDFs available, and 14 are pending PDF acquisition. This systematic approach, guided by our predefined inclusion and exclusion criteria, progressively narrowed down the selection to the most pertinent studies.

4.4 Snowballing

To complement the automated search and ensure no critical papers were missed, a snowballing technique was applied. This involved examining the reference lists of included studies and identifying papers that met our selection criteria, further enriching our dataset.

4.5 Research Questions

Our systematic literature review is guided by the following research questions:

- (1) What is the state of published literature on steganographic techniques that leverage large language models (LLMs)?
- (2) In which applications are steganographic techniques with LLMs being explored?
- (3) What metrics and evaluation methods are used to assess the performance of steganographic techniques in LLMs, focusing on factors like capacity, security, and contextual compatibility?
- (4) How are external knowledge sources (semantic resources) integrated into steganographic techniques with LLMs to enhance capacity or contextual relevance?
- (5) What are the limitations and trade-offs associated with current steganographic techniques using LLMs, particularly concerning security, capacity, and contextual compatibility?
- (6) What are the potential future research directions in steganography with LLMs, considering emerging trends and identified gaps in the literature?

5 DATA EXTRACTION AND CLASSIFICATION

This section outlines the methodology employed for extracting and classifying data from the selected primary studies. A structured approach was adopted to ensure consistency and accuracy in data collection, facilitating a comprehensive analysis of the literature.

5.1 Data Extraction Form (DEF) Content

A Data Extraction Form (DEF) was developed to systematically collect relevant information from each primary study. The DEF was designed to capture key details necessary to answer the research questions, including:

- **Title:** The title of the paper or resource.
- **Type:** State "Steganography" or "Watermarking."
- **Model Input:** Describe the input data format and its key characteristics for the model.
- **Model Output:** Describe the output format and its key characteristics of the model.
- **Categories:** Describe the approach using exactly three terms.
- **LLM (Large Language Model):** Specify the particular LLM used, if applicable.
- **Datasets Used:** List all datasets employed, including their sizes and any relevant details.
- **Main Strengths:** Identify and describe the primary strengths of the approach or model.
- **Main Weaknesses:** Identify and describe the primary weaknesses or limitations of the approach or model.
- **Evaluation Metrics and Steganalysis Models Used:** Detail the metrics used for evaluation and any steganalysis models applied.
- **Results (Best Metrics):** Present only the best numerical results for each reported metric.
- **Code Availability:** Indicate "Yes" or "No," and provide a link if available.
- **Embedding Process:** Provide a high-level, concise description of the data embedding process within the pipeline (e.g.,

"Word2Vec for synonyms, POS tagging for syntax, Universal Sentence Encoder for scoring"). Do not include method names.

- **Context Awareness:** State explicitly whether the method is "Explicit" (cares about the channel explicitly), "Implicit" (uses channel elements implicitly), or "No" (has no room for context). Context refers to the channel (e.g., chat, text) where the resultant (stego-text/marked text) is sent.
- **Categorical Context:** Describe with one keyword (e.g., "Social Media," "Formal Document").
- **Context Representation:** Explain how context is represented (e.g., "Text," "Pretext," "Graph," "Vector").
- **Context Usage in Method:** Detail how context is utilized within the method (free text).

5.2 Data Classification

Following data extraction, studies were classified based on predefined categories derived from our research questions. This classification aimed to group similar studies and identify trends, patterns, and gaps in the existing literature, providing a structured overview of the research landscape.

5.3 Presentation of Results

The results of the data synthesis are presented in a structured manner, often utilizing tables, figures, and descriptive statistics to summarize key findings. This includes an overview of publication trends, distribution of studies across different categories, and the prevalence of various approaches and techniques.

5.4 Discussion in Relation to Research Questions

Each research question is addressed individually, with a detailed discussion of the synthesized data. This involves interpreting the findings, highlighting significant observations, and drawing conclusions based on the evidence gathered from the primary studies. The discussion also identifies areas where further research is needed and potential future directions.

6 RESULTS AND DISCUSSION

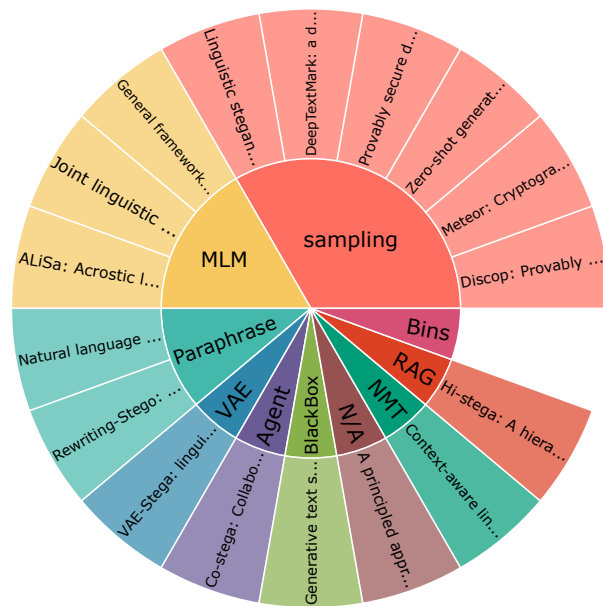


Figure 1: Sunburst Chart of LLM Approaches

Table 1: Summary of Results from Reviewed Papers

Paper	Result
VAE-Stega: linguistic steganography based on variational auto-encoder [15]	PPL: 28.879, ΔMP: 0.242, KLD: 3.302, JSD: 10.411, Acc: 0.600, R: 0.616
General framework for reversible data hiding in texts based on masked language modeling [18]	BPW=0.5335 F1=0.9402 PPL=134.2199
Co-stega: Collaborative linguistic steganography for the low capacity challenge in social media [7]	SR1: 60.87%, SR2: 98.55%, Gen. Capacity: 44.91 bits, Entropy: 49.21 bits, BPW: 2.31, PPL: 16.75, SimCSE: 0.69
Joint linguistic steganography with BERT masked language model and graph attention network [2]	PPL=13.917 KLD=2.904 SIM=0.812 ER=0.365 (BN=2) Best Acc=0.575 (BERT classifier) FLOPs=1.834G
Discop: Provably secure steganography in practice based on "distribution copies" [3]	p=1.00 Total Time (seconds)=362.63 Ave Time ↓ (seconds/bit)=6.29E-03 Ave KLD ↓ (bits/token)=0 Max KLD ↓ (bits/token)=0 Capacity (bits/token)=5.76 Entropy (bits/token)=6.08 Utilization ↑=0.95 Text Generation (FCN): 50.10%. Text Generation (R-BiLSTM-C): 50.45%. Text Generation (BiLSTM-Dense): 49.95%
Generative text steganography with large language model [14]	Length: 13.333 (words). BPW: 5.93 bpw PPL: 165.76. Semantic Similarity (SS): 0.5881 LS-CNN Acc: 51.55%. BiLSTM-Dense Acc: 49.20%. Bert-FT Acc: 50.00%. KLD (Log, lower is better): 2.02 .
Meteor: Cryptographically secure steganography for realistic distributions [5]	GPT-2: 3.09 bits/token
Zero-shot generative linguistic steganography [8]	PPL: 8.81. JSDfull: 17.90 ($\times 10^{-2}$). JSDhalf: 16.86 ($\times 10^{-2}$). JSDzero: 13.40 ($\times 10^{-2}$) TS-BiRNN: 80.29%. R-BiLSTM-C: 84.34%. BERT-C: 89.61%
Provably secure disambiguating neural linguistic steganography [10]	Total Error: 0%, Ave KLD: 0, Max KLD: 0, Ave PPL: 3.19 (EN), 7.49 (ZH), Capacity: 1.03–3.05 bits/token, Utilization: 0.66–0.74, Ave Time: 4μs/bit
A principled approach to natural language watermarking [4]	Bit acc: 0.994 (K=None), 1.000 (DAE), 0.978 (Adaptive+K=S); Meteor Drop: 0.057; SBERT ↑: 1.227; Ownership Rate: 1.0 (no attack), 0.978 (adaptive+K=S)
Context-aware linguistic steganography model based on neural machine translation [1]	BLEU: 30.5, PPL: 22.5, ER: 0.29, KL: 0.02, SIM: 0.86, Stego detection 16%
DeepTextMark: a deep learning-driven text watermarking approach for identifying large language model generated text [9]	100% accuracy (multi-synonym, 10-sentence), mSMS: 0.9892, TPR: 0.83, FNR: 0.17, Detection: 0.00188s, Insertion: 0.27931s
Hi-stega: A hierarchical linguistic steganography framework combining retrieval and generation [13]	ppl: 109.60, MAUVE: 0.2051, ER2: 10.42, Δ(cosine): 0.0088, Δ(simcse): 0.0191
Linguistic steganography: From symbolic space to semantic space [17]	Classifier Accuracy: 0.9880; Loop Count: 1.0160; PPL: 13.9565; Anti-Steganalysis Accuracy: 0.5
Natural language steganography by chatgpt [12]	[Not specified]
Natural language watermarking via paraphraser-based lexical substitution [11]	LS07 P@1: 58.3, GAP: 65.1; CoInCo P@1: 62.6, GAP: 60.7; Text Recoverability: 88–90%
Rewriting-Stego: generating natural and controllable steganographic text with pre-trained language model [6]	BPTS: 4.0, BPTC+S: 4.0, PPL: 62.1, Mean: 44.4, Variance: 2.1e04, Acc: 8.9%
ALiSa: Acrostic linguistic steganography based on BERT and Gibbs sampling [16]	PPL: Natural = 13.91, ALiSa = 14.85; LS-RNN/LS-BERT Acc & F1 = 0.50; Outperforms GPT-AC/ADG in all cases

Table 2: Models and Datasets Used in Reviewed Papers

Paper	Llm	Dataset
VAE-Stega: linguistic steganography based on variational auto-encoder [15]	BERTBASE (BERT-LSTM) (LSTM-LSTM) model was trained from scratch	Twitter (2.6M sentences) IMDB (1.2M sentences) preprocessed
General framework for reversible data hiding in texts based on masked language modeling [18]	BERTBase	BookCorpus
Co-stega: Collaborative linguistic steganography for the low capacity challenge in social media [7]	Llama-2-7B-chat, GPT-2 (fine-tuned), Llama-2-13B	Tweet dataset (for GPT-2 fine-tuning), Twitter (real-time testing)
Joint linguistic steganography with BERT masked language model and graph attention network [2]	LSTM + attention for temporal context. GAT for spatial token relationships. BERT MLM for deep semantic context in substitution.	OPUS
Discop: Provably secure steganography in practice based on "distribution copies" [3]	GPT-2	IMDB
Generative text steganography with large language model [14]	Any	[Not specified]
Meteor: Cryptographically secure steganography for realistic distributions [5]	GPT-2	Hutter Prize, HTTP GET requests
Zero-shot generative linguistic steganography [8]	LLaMA2-Chat-7B (as the stegotext generator / QA model). GPT-2 (for NLS baseline and JSD evaluation)	IMDB, Twitter
Provably secure disambiguating neural linguistic steganography [10]	LLaMA2-7b (English), Baichuan2-7b (Chinese)	IMDb dataset (100 texts/sample, 3 English sentences + Chinese translations)
A principled approach to natural language watermarking [4]	Transformer-based encoder/decoder; BERT for distillation	Web Transformer 2
Context-aware linguistic steganography model based on neural machine translation [1]	BERT (encoder), LSTM (decoder)	WMT18 News Commentary (train/test), Yang et al. bits, Doc2Vec, 5,000 stego pairs (8:1:1 split)
DeepTextMark: a deep learning-driven text watermarking approach for identifying large language model generated text [9]	Model-independent; tested with OPT-2.7B	Dolly ChatGPT (train/validate), C4 (test), robustness & sentence-level test sets
Hi-stega: A hierarchical linguistic steganography framework combining retrieval and generation [13]	GPT-2	Yahoo! News (titles, bodies, comments); 2,400 titles used
Linguistic steganography: From symbolic space to semantic space [17]	CTRL (generation), BERT (semantic classifier)	5,000 CTRL-generated texts per semanteme (n = 2–16); 1,000 user-generated texts for anti-steganalysis
Natural language steganography by chatgpt [12]	[Not specified]	Custom word sets for specific topics (e.g., 16×10-word sets for music reviews)
Natural language watermarking via paraphraser-based lexical substitution [11]	Transformer (Paraphraser), BART (BARTScore), BERT (BLEURT, comparisons)	ParaBank2, LS07, CoInCo, Novels, WikiText-2, IMDB, NgNews
Rewriting-Stego: generating natural and controllable steganographic text with pre-trained language model [6]	BART (bart-base2)	Movie, News, Tweet
ALiSa: Acrostic linguistic steganography based on BERT and Gibbs sampling [16]	BERT (Google's BERTBase, Uncased)	BookCorpus (10,000 natural texts for evaluation)

6.1 State of Published Literature on LLM-based Steganography

This section summarizes the main findings from the systematic literature review, focusing on the characteristics and performance of various LLM-based linguistic steganography and watermarking models.

Large Language Models (LLMs) have emerged as a significant development in the field of natural language processing, profoundly impacting text generation and related applications like steganography and watermarking. Our review identified several key LLM-based steganography models, each with unique approaches, strengths, and performance metrics.

LLMs are **generative models** that can **approximate complex distributions like text-based communication**. They represent the best-known technique for this task. These models operate by taking context and parameters to output an explicit probability distribution over the next token (e.g., a character or a word). The next token is typically sampled randomly from this distribution, and the process repeats to generate output of a desired length.

The **quality of content generated by generative models is impressive** and continues to improve. This has led to LLMs blurring the boundary of high-quality text generation between humans and machines.

6.2 Applications of LLM-based Steganographic Techniques

LLMs are considered **favorable for generative text steganography** due to their ability to generate high-quality text. Researchers propose using generative models as steganographic samplers to embed messages into realistic communication distributions, such as text. This approach marks a departure from prior steganographic work, motivated by the public availability of high-quality models and significant efficiency gains.

LLMs like **GPT-2, LLaMA, and Baichuan2** are commonly used as basic generative models for steganography. Existing methods often utilize a language model and steganographic mapping, where secret messages are embedded by establishing a mapping between binary bits and the sampling probability of words within the training vocabulary.

New approaches, such as **LLM-Stega**, explore **black-box generative text steganography using the user interfaces (UIs) of LLMs**, thereby circumventing the requirement to access internal sampling distributions. This method constructs a keyword set and employs an encrypted steganographic mapping for embedding, proposing an optimization mechanism based on reject sampling for accurate extraction and rich semantics.

Another framework, **Co-Stega**, leverages LLMs to address the challenge of low capacity in social media by expanding the text space for hiding messages (through context retrieval) and **increasing the generated text's entropy via specific prompts** to enhance embedding capacity. This approach also aims to maintain text quality, fluency, and relevance.

The concept of **zero-shot linguistic steganography** with LLMs utilizes in-context learning, where samples of covertext are used as context to generate more intelligible stegotext using a question-answer (QA) paradigm.

LLMs are also employed in approaches like **ALiSa**, which directly conceals token-level secret messages in seemingly natural steganographic text generated by off-the-shelf BERT models equipped with Gibbs sampling.

6.3 Evaluation Metrics and Methods for LLM-based Steganography

6.3.1 Imperceptibility Metrics. Perceptual metrics include PPL (Perplexity), Distinct-n, MAUVE, and human evaluation. Statistical metrics include KLD (Kullback-Leibler Divergence), JSD (Jensen-Shannon Divergence), anti-steganalysis accuracy, and semantic similarity.

6.3.2 Embedding Capacity Metrics. Metrics include bits per token/word and embedding rate.

6.4 Integration of External Knowledge Sources

The increasing popularity of deep generative models has made it feasible for provably secure steganography to be applied in real-world scenarios, as they fulfill requirements for perfect samplers and explicit data distributions.

Some approaches leverage LLMs to address the challenge of low capacity in social media by expanding the text space for hiding messages through context retrieval. This integration of external knowledge enhances both the capacity and contextual relevance of the steganographic techniques.

6.5 Limitations and Trade-offs in Current LLM-based Steganography

6.5.1 Perceptual vs. Statistical Imperceptibility (Psic Effect). The Psic Effect represents a fundamental trade-off where improving perceptual quality can reduce statistical security, and vice versa. This fundamental trade-off presents a significant challenge in the field.

6.5.2 Low Embedding Capacity. Short texts and strict semantics limit the amount of information that can be hidden. This is a particular challenge in applications where the cover text must appear natural and contextually appropriate.

6.5.3 Lack of Semantic Control and Contextual Consistency. Ensuring generated text matches intended meaning and context is difficult. LLMs may introduce unpredictability, bias, or leak information.

6.5.4 Segmentation Ambiguity. Subword tokenization in LLMs can create ambiguity in message extraction, as the same text can be tokenized differently depending on context.

6.5.5 White-box vs. Black-box Access. Traditional "white-box" methods necessitate sharing the exact language model and training vocabulary, which limits fluency, logic, and diversity compared to natural texts generated by LLMs. These methods also inevitably alter the sampling probability distribution, thereby posing security risks.

6.5.6 Other Challenges. Additional challenges include computational overhead, data integrity/reversibility issues, and ethical concerns such as biases, discrimination, and potential for generating

Table 3: Context-Related Fields in Reviewed Papers

Paper	Context Aware	Categ Context	Representation Context
VAE-Stega: linguistic steganography based on variational auto-encoder [15]	non-explicit	pre-text	text
General framework for reversible data hiding in texts based on masked language modeling [18]	non-explicit	pre-text	text
Co-stega: Collaborative linguistic steganography for the low capacity challenge in social media [7]	explicit	Social Media	text
Joint linguistic steganography with BERT masked language model and graph attention network [2]	explicit	pre-text	text
Discop: Provably secure steganography in practice based on "distribution copies" [3]	non-explicit	tuning + pretext	text
Generative text steganography with large language model [14]	explicit	[Not specified]	[Not specified]
Meteor: Cryptographically secure steganography for realistic distributions [5]	non-explicit	tuning + pretext	text
Zero-shot generative linguistic steganography [8]	explicit	zero-shot + prompt	text
Provably secure disambiguating neural linguistic steganography [10]	non-explicit	pretext	text
A principled approach to natural language watermarking [4]	Yes; semantic-level embedding; synonym substitution using BERT	Yes; watermark message assigned categorical label (e.g., 4-bit \rightarrow 1-of-16)	Yes; semantic embeddings via transformer encoder and BERT; SBERT distance as metric
Context-aware linguistic steganography model based on neural machine translation [1]	Yes	[Not specified]	GCF (global context), LMR (language model reference), Multi-head attention
DeepTextMark: a deep learning-driven text watermarking approach for identifying large language model generated text [9]	NO	[Not specified]	[Not specified]
Hi-stega: A hierarchical linguistic steganography framework combining retrieval and generation [13]	explicit	Social Media	Text
Linguistic steganography: From symbolic space to semantic space [17]	implicit	Text	Semanteme (α) as a vector in semantic space
Natural language steganography by chatgpt [12]	Explicit	Specific Genre/Topic Text	Text
Natural language watermarking via paraphraser-based lexical substitution [11]	Explicit	[Not specified]	text
Rewriting-Stego: generating natural and controllable steganographic text with pre-trained language model [6]	not Explicit	[Not specified]	[Not specified]
ALiSa: Acrostic linguistic steganography based on BERT and Gibbs sampling [16]	No	[Not specified]	[Not specified]

insulting content. There is also a lack of provable security and rigor in many NLP steganography works.

6.6 Future Research Directions

Based on the identified gaps and challenges, several promising future research directions emerge:

- **Improved Balance Between Perceptual and Statistical Imperceptibility:** Developing techniques that can maintain both high perceptual quality and statistical security.
- **Enhanced Embedding Capacity:** Exploring methods to increase the amount of information that can be hidden without compromising imperceptibility.
- **Better Semantic Control:** Advancing approaches that ensure generated steganographic text maintains intended meaning and contextual consistency.
- **Addressing Segmentation Ambiguity:** Developing robust techniques to handle the challenges posed by subword tokenization in LLMs.
- **Ethical Frameworks:** Establishing guidelines and frameworks for the ethical use of LLM-based steganography to prevent misuse.
- **Provable Security:** Advancing the theoretical foundations of LLM-based steganography to provide stronger security guarantees.
- **Efficient Computation:** Reducing the computational overhead associated with LLM-based steganography techniques.

The field of LLM-based steganography is rapidly evolving, with new models and techniques being developed to address these challenges and explore new possibilities.

7 MAIN FINDINGS

This section summarizes the key findings from our systematic literature review on LLM-based steganography techniques.

7.1 Overview of LLM-based Steganography

Large Language Models (LLMs) have revolutionized the field of linguistic steganography by providing high-quality text generation capabilities that can be leveraged for information hiding. Our review has identified several important trends and developments in this emerging field:

- LLMs like GPT-2, LLaMA, and Baichuan2 are increasingly being used as the foundation for steganographic techniques due to their ability to generate natural-sounding text.
- Both white-box approaches (with access to model internals) and black-box approaches (using only model interfaces) have been developed, each with distinct advantages and limitations.
- The field faces fundamental trade-offs between imperceptibility, capacity, and security that continue to drive research innovation.

7.2 Key Techniques and Approaches

Our analysis identified several innovative approaches to LLM-based steganography:

- **LLM-Stega:** A black-box approach that uses the user interfaces of LLMs without requiring access to internal sampling distributions.
- **Co-Stega:** Leverages LLMs to expand text space for hiding messages through context retrieval and increases text entropy via specific prompts.
- **Zero-shot linguistic steganography:** Utilizes in-context learning with a question-answer paradigm to generate more natural stegotext.
- **ALiSa:** Conceals token-level secret messages in natural-looking text generated by BERT models with Gibbs sampling.

7.3 Critical Challenges

Despite significant progress, several challenges remain in the field of LLM-based steganography:

- The Psic Effect: A fundamental trade-off between perceptual quality and statistical security.
- Limited embedding capacity, particularly in short texts with strict semantic requirements.
- Difficulties in maintaining semantic control and contextual consistency in generated steganographic text.
- Segmentation ambiguity arising from subword tokenization in LLMs.
- Ethical concerns related to potential misuse, bias, and discrimination in generated content.

7.4 Future Outlook

Based on our analysis, we identify several promising directions for future research:

- Development of techniques that better balance perceptual quality and statistical security.
- Methods to increase embedding capacity without compromising imperceptibility.
- Approaches to improve semantic control and contextual consistency in generated text.
- Frameworks for ethical use of LLM-based steganography.
- Advancement of theoretical foundations to provide stronger security guarantees.

The rapid evolution of LLMs presents both opportunities and challenges for the field of steganography, making it an exciting area for continued research and innovation.

8 CONCLUSION

Summarize the main findings and takeaways of the study.

REFERENCES

- [1] Changhao Ding, Zhangjie Fu, Zhongliang Yang, Qi Yu, Daqiu Li, and Yongfeng Huang. 2023. Context-aware linguistic steganography model based on neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2023), 868–878.
- [2] Changhao Ding, Zhangjie Fu, Qi Yu, Fan Wang, and Xianyi Chen. 2023. Joint linguistic steganography with BERT masked language model and graph attention network. *IEEE Transactions on Cognitive and Developmental Systems* 16, 2 (2023), 772–781.
- [3] Jinyang Ding, Kejiang Chen, Yaofei Wang, Na Zhao, Weiming Zhang, and Nenghai Yu. 2023. Discop: Provably secure steganography in practice based on "distribution copies". In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2238–2255.

- [4] Zhe Ji, Qiansiqi Hu, Yicheng Zheng, Liyao Xiang, and Xinbing Wang. 2024. A principled approach to natural language watermarking. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2908–2916.
- [5] Gabriel Kaptchuk, Tushar M Jois, Matthew Green, and Aviel D Rubin. 2021. Meteor: Cryptographically secure steganography for realistic distributions. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 1529–1548.
- [6] Fanxiao Li, Sixing Wu, Jiong Yu, Shuoxin Wang, BingBing Song, Renyang Liu, Haoseng Lai, and Wei Zhou. 2023. Rewriting-Stego: generating natural and controllable steganographic text with pre-trained language model. In *International Conference on Database Systems for Advanced Applications*. Springer, 617–626.
- [7] Guorui Liao, Jinshuai Yang, Kaiyi Pang, and Yongfeng Huang. 2024. Co-stega: Collaborative linguistic steganography for the low capacity challenge in social media. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*. 7–12.
- [8] Ke Lin, Yiyang Luo, Zijian Zhang, and Ping Luo. 2024. Zero-shot generative linguistic steganography. *arXiv preprint arXiv:2403.10856* (2024).
- [9] Travis Munyer, Abdullah All Tanvir, Arjon Das, and Xin Zhong. 2024. Deep-TextMark: a deep learning-driven text watermarking approach for identifying large language model generated text. *Ieee Access* 12 (2024), 40508–40520.
- [10] Yang Qi, Kejiang Chen, Kai Zeng, Weiming Zhang, and Nenghai Yu. 2024. Provably secure disambiguating neural linguistic steganography. *IEEE Transactions on Dependable and Secure Computing* (2024).
- [11] Jipeng Qiang, Shiyu Zhu, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. Natural language watermarking via paraphraser-based lexical substitution. *Artificial Intelligence* 317 (2023), 103859.
- [12] Martin Steinebach. 2024. Natural language steganography by chatgpt. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*. 1–9.
- [13] Huili Wang, Zhongliang Yang, Jinshuai Yang, Yue Gao, and Yongfeng Huang. 2023. Hi-stega: A hierarchical linguistic steganography framework combining retrieval and generation. In *International Conference on Neural Information Processing*. Springer, 41–54.
- [14] Jiaxuan Wu, Zhengxian Wu, Yiming Xue, Juan Wen, and Wanli Peng. 2024. Generative text steganography with large language model. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 10345–10353.
- [15] Zhong-Liang Yang, Si-Yu Zhang, Yu-Ting Hu, Zhi-Wen Hu, and Yong-Feng Huang. 2020. VAE-Stega: linguistic steganography based on variational auto-encoder. *IEEE Transactions on Information Forensics and Security* 16 (2020), 880–895.
- [16] Biao Yi, Hanzhou Wu, Guorui Feng, and Xinpeng Zhang. 2022. ALiSa: Acros-tic linguistic steganography based on BERT and Gibbs sampling. *IEEE Signal Processing Letters* 29 (2022), 687–691.
- [17] Siyu Zhang, Zhongliang Yang, Jinshuai Yang, and Yongfeng Huang. 2020. Linguistic steganography: From symbolic space to semantic space. *IEEE Signal Processing Letters* 28 (2020), 11–15.
- [18] Xiaoyan Zheng, Yurun Fang, and Hanzhou Wu. 2022. General framework for reversible data hiding in texts based on masked language modeling. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–6.