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

NASOUH ALOLABI, Higher Institute for Applied Sciences and Technology, Syria
RIAD SONBOL, Higher Institute for Applied Sciences and Technology, Syria

This systematic literature review examines the transformative impact of Large Language Models (LLMs) on linguistic steganography. Through comprehensive analysis of 18 primary studies and 14 additional papers, the research demonstrates that LLM-based approaches significantly enhance imperceptibility (achieving PPL scores of 3-8 for white-box methods), embedding capacity (up to 5.98 bits per token), and naturalness in cover text generation, addressing traditional limitations of low embedding capacity and cognitive imperceptibility. The 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 advancements in generative text steganography.

Additional Key Words and Phrases: Systematic Literature Review, Linguistic Steganography, Large Language Models, LLMs, Natural Language Processing, NLP, Black-box Steganography, Context Retrieval, Generative Text Steganography, Imperceptibility

Preprint Notice: This is a preprint version of our systematic literature review, last updated on August 12, 2025. The work is currently under review for publication.

1 INTRODUCTION

Linguistic steganography, the practice of concealing information within natural language text, has long been regarded as one of the most challenging areas of covert communication due to the low redundancy [42] [16], semantic rigidity, and statistical sensitivity of language. Traditional methods —such as synonym substitution, syntactic transformations, or rule-based embedding— often suffer from limited capacity and detectability [13], making them inadequate against modern steganalysis. The emergence of large language models (LLMs), however, has profoundly transformed this landscape by enabling the generation of coherent, context-aware, and statistically natural covertexts [40], thereby providing a foundation for high-capacity and imperceptible covert communication. The field has seen the emergence of various LLM-based steganography paradigms: generative methods that directly create stego texts [42][45][10][38], rewriting-based methods that rephrase existing cover texts [18], black-box approaches that utilize LLM user interfaces or APIs without needing access to internal model parameters [38][34], zero-shot methods that leverage in-context learning in contrast to fine tuning with LLMs to generate intelligible stego text [21], collaborative frameworks that exploit contextual relevance within social media or combine retrieval and generation strategies to expand embedding space and enhance entropy [20][37], provably secure methods that focus on mathematically rigorous security definitions, achieving indistinguishability from honest model output [16][10]. While LLMs offer significant advantages, challenges like the "Psic Effect" (a trade-off between text quality and statistical imperceptibility) [42], computational overhead, and segmentation ambiguity still present areas for ongoing research. This paper presents a systematic literature review that synthesizes recent advances in LLM-based linguistic steganography, identifies unresolved challenges, and highlights future research directions.

Authors' addresses: Nasouh AlOlabi, Higher Institute for Applied Sciences and Technology, Damascus, Syria; Riad Sonbol, Higher Institute for Applied Sciences and Technology, Damascus, Syria.

Previous reviews on text steganography, such as the one by Majeed et al. (2021) [23], primarily focus on older techniques and were published before the widespread adoption of Large Language Model (LLM)-based approaches. While the more recent review by Setiadi et al. (2025) [31] acknowledges that the field of linguistic steganography "has been revitalized by large language models (LLMs)" and specifically examines recent AI-powered steganography methods from the last three years (post-2021), detailing techniques that utilize models like GPT-2 [29], GPT-3 [1], LLaMA2 [2], and Baichuan2 [39], it is important to note that the Setiadi et al. (2025) review is not a systematic literature review. It's a "concise and critical examination" rather than an exhaustive survey, it does not include all relevant papers published between 2021 and 2025. Consequently, despite the advancements discussed, a notable gap persists for a comprehensive systematic literature review that fully summarizes how large-scale transformers have reshaped text steganography. This is in contrast to earlier surveys that predominantly identified classical approaches such as synonym replacement, spacing, and Huffman coding, which predated the LLM revolution [23].

Furthermore, the field faces significant challenges in evaluation standardization that compound the need for systematic analysis. While core metrics like embedding rate (ER) [6] , Kullback-Leibler divergence (KLD) [17] , and perplexity (PPL) [14] are consistently used across studies, their inconsistent application hinders meaningful cross-method comparisons. For instance, PPL calculations vary depending on the underlying language model used (GPT-2, LLaMA, etc.) and the generated text length, KLD measurements differ based on the reference datasets (normal text) employed, and ER reporting lacks uniformity with some studies measuring bits per token while others use bits per word. This inconsistency is compounded by the use of heterogeneous datasets across studies, ranging from IMDb [22] and BookCorpus [48] to specialized corpora like News-Commentary-v13 [define/reference needed] and HC3 [define/reference needed]. Unlike image steganography, which benefits from standardized visual quality metrics such as PSNR [define/reference needed] and SSIM [define/reference needed], linguistic steganography [define/reference needed] lacks unified evaluation protocols, making objective performance comparisons challenging and potentially misleading [citation needed].

This systematic review fills these gaps by meticulously identifying and synthesizing recent primary literature that leverages LLMs for textual steganography, particularly from the last two years when LLMs like GPT-3/4 [citation/reference needed] and open models became widely available [citation/reference needed]. The timing is well-justified by the significant surge in publications and novel ideas since 2023 [citation/reference needed], with approximately 70% of recent studies using open-source LLMs like GPT-2 [citation/reference needed], LLaMA2 [citation/reference needed], and LLaMA3 [citation/reference needed]. The importance of this review is underscored by the transformative impact of LLMs on secure communication [citation/reference needed], marking a paradigm shift toward context-aware, generative systems that prioritize imperceptibility, embedding capacity, and naturalness [citation/reference needed]. LLM-based steganography offers striking gains in classic metrics like capacity and imperceptibility [citation/reference needed]; for instance, reviewed studies report that advanced white-box LLM samplers can achieve perplexities as low as 3-8 (on GPT-2 models) while embedding up to approximately 5.98 bits per token [citation/reference needed], far exceeding pre-LLM schemes [citation/reference needed]. This enables secure clandestine messaging in environments where classical steganography was too limited or suspicious [citation/reference needed].

The rest of this paper follows a standard SLR structure. Section 2 provides background on steganography and LLMs, defining key concepts such as imperceptibility. Section 3 describes the scope and research questions. Section 4 details the literature search and selection methodology. Sections 5 and 6 present the data extraction process and classification of the selected studies. Section 7 reports the results organized by research question, summarizing state-of-the-art techniques, application domains, evaluation metrics, attack models, and the role of external knowledge sources. Finally, Section 8 synthesizes the main findings and discusses trends, and Section 9 concludes by outlining open problems and future research directions.

2 BACKGROUND

Information security systems broadly encompass **encryption**, **privacy**, and **concealment**, the last of which—known as **steganography**—is the focus of this review. While encryption and privacy protect message content, they do not conceal the existence of communication, which may itself arouse suspicion. Steganography instead prioritizes **imperceptibility**: embedding information into ordinary carriers (e.g., images or text) so that hidden messages remain unnoticed.

Text is a particularly challenging carrier due to its low redundancy and strict semantic constraints. The classical "Prisoners' Problem" [33] illustrates the goal: two parties, Alice and Bob, must exchange hidden information without alerting a watchful adversary.

Textual steganography methods are typically divided into **format-based** approaches, which exploit layout or structural features, and **content-based** approaches, which modify linguistic form. Within the latter, early techniques such as **synonym substitution** embed bits by altering lexical choices, but suffer from low capacity and high detectability. More formally, **linguistic steganography** refers to concealing information in natural language by modifying or generating text while preserving fluency and meaning [11].

Traditional linguistic approaches offer limited embedding capacity and often leave statistical artifacts. Advances in deep learning and **Large Language Models (LLMs)** now enable generative methods that achieve higher text quality and more secure embedding. Evaluating such systems requires several dimensions of imperceptibility: **perceptual** (human naturalness), **statistical** (distributional similarity to natural text), and **cognitive** (semantic and contextual fidelity) [8].

A deeper theoretical perspective introduces **channel entropy**, which quantifies the information-carrying capacity of a given communication channel. Entropy sets the upper bound for embedding rates: higher entropy allows more hidden information without detection, while lower entropy restricts capacity. Achieving this bound securely requires **perfect samplers**, which can generate text indistinguishable from genuine distributional samples. These concepts underpin the design of provably secure steganographic systems.

However, LLMs [32] introduce new challenges. Their tendency toward **hallucinations** can create detectable artifacts, highlighting the **Psic Effect** (Perceptual-Statistical Imperceptibility Conflict) [42], where optimizing for perceptual fluency may undermine statistical security. Model access further shapes practical steganography: with **black-box access** (e.g., commercial APIs), developers gain scalability and ease of use but face limited control and reduced transparency. In contrast, **white-box access** enables fine-grained control over parameters and sampling, supporting stronger security guarantees, but requires costly resources and raises deployment barriers. This trade-off is central to evaluating the robustness and applicability of modern linguistic steganography.

3 STEGANOGRAPHY AND LARGE LANGUAGE MODELS

3.1 Capabilities and Approximating Natural Communication

Large Language Models (LLMs) are autoregressive, generative systems based on the Transformer architecture [36] that approximate high-dimensional distributions over natural-language sequences [16][30]. Given a prefix, an LLM emits a probability vector over the vocabulary; the next token is sampled from this vector and appended to the prefix, and the process repeats until a stopping criterion is met. During pre-training, billions of parameters are tuned on large web corpora so that the model's predictive distribution converges to the empirical distribution of the data [4]. As a consequence, modern LLMs routinely produce text whose fluency, coherence and style are indistinguishable from human writing [5]. The learned latent representations capture stylistic and semantic regularities that generalize across domains, enabling applications requiring nuanced linguistic mimicry [46].

3.2 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** [30], **LLaMA** [35], and **Baichuan2** [41] 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 [38].

New approaches, such as **LLM-Stega** [38], explore **black-box generative text steganography using the user interfaces (UIs) of LLMs**. This circumvents the requirement to access internal sampling distributions. The method constructs a keyword set and employs an encrypted steganographic mapping for embedding. It proposes an optimization mechanism based on reject sampling for accurate extraction and rich semantics [38].

Another framework, **Co-Stega**, leverages LLMs to address the challenge of low capacity in social media. It expands the text space for hiding messages through context retrieval and **increases the generated text's entropy via specific prompts** to enhance embedding capacity. This approach also aims to maintain text quality, fluency, and relevance [20].

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 [21]. 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 [7] models equipped with Gibbs sampling [43].

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 (see Section **??**) [10, 16, 27].

3.3 LLM-Based Steganography Models

3.3.1 Evaluation Metrics.

Imperceptibility Metrics. Perceptual metrics include PPL [12], Distinct-n [19], MAUVE [26], and human evaluation. Statistical metrics include KLD, JSD, anti-steganalysis accuracy, and semantic similarity [25].

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

3.4 Challenges and Limitations in Steganography with LLMs

3.4.1 Perceptual vs. Statistical Imperceptibility (Psic Effect). The **Psic Effect** [42] represents a fundamental trade-off in steganographic systems.

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

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

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

3.4.5 *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 [38]. 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 [10, 16, 38, 42].

Another significant hurdle is **ensuring both the quality and imperceptibility of the generated text**, encompassing perceptual, statistical, and cognitive imperceptibility [8]. 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" [8, 42]. 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 [8, 37].

Channel entropy requirements and variability also pose a considerable challenge. Traditional universal steganographic schemes often demand consistent channel entropy, which is rarely maintained in real-world natural language communication. Moments of low or zero entropy can cause protocols to fail or require extraordinarily long steganographic texts. The Psic Effect highlights this dilemma in balancing quality and detectability.

Furthermore, **segmentation ambiguity** introduced by subword-based language models presents a critical issue for provably secure linguistic steganography. When a sender detokenizes generated subword sequences into continuous text, the receiver might retokenize it differently, leading to decoding errors [27].

Additional limitations include:

- **Computational Overhead**: LLMs incur 3-5 times higher computational cost than prior methods [21].
- **Data Integrity and Reversibility**: Some methods cannot perfectly recover the original cover text after message extraction [28, 47].
- **Ethical Concerns**: Pre-trained LLMs may introduce biases, discrimination, or inappropriate content [3, 21].
- **Provable Security**: Many NLP steganography works lack rigorous security analyses and fail to meet formal cryptographic definitions [16].
- **Hallucinations**: LLMs can generate factually incorrect or contextually inappropriate content, leading to embedding errors [12].

- **Channel Entropy Limitations**: Short, context-dependent texts have lower entropy, limiting hiding capacity [20].

4 LITERATURE REVIEW METHODOLOGY

4.1 Research questions

The research questions addressed in this systematic literature review are:

- 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?

4.2 Search query string

The following search query string was employed for the initial literature search:

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

4.3 Study selection and quality assessment

The following inclusion and exclusion criteria were established for study selection:

4.3.1 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.

4.3.2 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).

4.4 Bibliometric analysis

Briefly note if snowballing was used for additional sources.

4.5 Threats to Validity

While this systematic literature review (SLR) adheres to established guidelines such as PRISMA to ensure methodological rigor, several potential threats to validity must be acknowledged. These threats primarily relate to the comprehensiveness of the literature search, selection biases, and practical constraints in data acquisition.

First, the search strategy may introduce publication and selection biases. The query string was limited to English-language publications from 2018 onward, potentially excluding relevant non-English studies or foundational pre-2018 works on linguistic steganography that predate widespread LLM adoption. Although LLMs emerged prominently around 2018 with models such as BERT, this cutoff might overlook influential earlier contributions that inform current techniques. Additionally, the selected databases (ACM Digital Library, IEEE Digital Library, Science@Direct, Scopus, and Springer Link) provide broad coverage but may miss papers in other repositories, including arXiv, Google Scholar, or domain-specific journals. The search terms, while comprehensive, could overlook synonyms or emerging variants (e.g., "textual watermarking" without explicit LLM mentions), despite efforts to include related phrases such as "Information Hiding."

Second, biases in study selection and quality assessment could affect the review's internal validity. The inclusion criteria focused on peer-reviewed sources, which enhances reliability but may introduce publication bias by favoring positive or novel results over negative findings or gray literature. No formal risk-of-bias tool (e.g., ROBIS) was applied beyond basic relevance checks, potentially allowing lower-quality studies to influence findings. To mitigate this, multi-stage filtering with title, abstract, and full-text reviews was employed, and snowballing was used to identify additional references, though it primarily yielded older non-LLM works.

Third, practical limitations pose threats to completeness. As noted in Section 4.3, 14 papers remained pending PDF acquisition at the time of analysis, which could lead to incomplete coverage if these contain critical insights. This issue was addressed by prioritizing accessible studies and planning follow-up acquisition, but it highlights retrieval challenges in SLR processes.

Overall, these threats were minimized through transparent documentation of the methodology, adherence to PRISMA reporting standards, and supplementary snowballing. Future updates to this review could expand database coverage and incorporate automated tools for bias assessment to further enhance validity.

5 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.

5.1 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.

5.2 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.

5.3 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.

5.4 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. Notably, all references identified through snowballing were to papers employing older steganographic techniques that do not explicitly mention the term "LLM" but utilize similar methodological approaches to those found in contemporary LLM-based steganography.

5.5 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?

6 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.

6.1 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 for addressing 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).

6.2 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.

6.3 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.

6.4 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

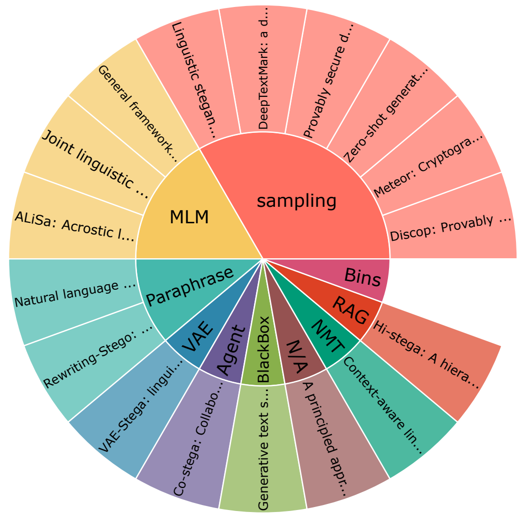

Fig. 1. Sunburst Chart of LLM Approaches

from the primary studies. The discussion also identifies areas where further research is needed and potential future directions.

Table 1. Summary of Results from Reviewed Papers

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
VAE-Stega: linguistic steganography based on va... [42]	BERTBASE (BERT-LSTM) (LSTM-LSTM) model was trained from scratch	Twitter (2.6M sentences) IMDB (1.2M sentences) preprocessed	PPL: 28.879, ΔMP: 0.242, KLD: 3.302, JSD: 10.411, Acc: 0.600, R: 0.616	non-explicit	pre-text	text
General framework for reversible data hiding in... [47]	BERTBase	BookCorpus	BPW=0.5335 F1=0.9402 PPL=134.2199	non-explicit	pre-text	text

Continued on next page

Table 1 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
Co-stega: Collaborative linguistic stegano-graph... [20]	Llama-2-7B-chat, GPT-2 (fine-tuned), Llama-2-13B	Tweet dataset (for GPT-2 fine-tuning), Twitter (real-time testing)	SR1: 60.87%, SR2: 98.55%, Gen. Capacity: 44.91 bits, Entropy: 49.21 bits, BPW: 2.31, PPL: 16.75, SimCSE: 0.69	explicit	Social Media	text
Joint linguistic steganography with BERT masked... [9]	LSTM + attention for temporal context. GAT for spatial token relationships. BERT MLM for deep semantic context in substitution.	OPUS	PPL=13.917 KLD=2.904 SIM=0.812 ER=0.365 (BN=2) Best Acc=0.575 (BERT classifier) FLOPs=1.834G	explicit	pre-text	text
Discop: Provably secure steganography in practi...	GPT-2	IMDB	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 E...	non-explicit	tuning + pre-text	text
Generative text steganography with large langua... [38]	Any	[Not specified]	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...	explicit	[Not specified]	[Not specified]

Continued on next page

Table 1 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
Meteor: Cryptographically secure steganography ... [16]	GPT-2	Hutter Prize, HTTP GET requests	GPT-2: 3.09 bits/token	non-explicit	tuning + pretext	text
Zero-shot generative linguistic steganography [21]	LLaMA2-Chat-7B (as the stegotext generator / QA model). GPT-2 (for NLS baseline and JSD evaluation)	IMDB, Twitter	PPL: 8.81. JS-Dfull: 17.90 (x10[truncated]iicircum-2). JSDhalf: 16.86 (x10[truncated]iicircum-2). JSDzero: 13.40 (x10[truncated]iicircum-2) TS...	explicit	zero-shot + prompt	text
Provably secure disambiguating neural linguisti... [27]	LLaMA2-7b (English), Baichuan2-7b (Chinese)	IMDb dataset (100 texts/sample, 3 English sentences + Chinese translations)	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: [truncat...	non-explicit	pretext	text
A principled approach to natural language water... [15]	Transformer-based encoder/decoder; BERT for distillation	Web Transformer 2	Bit acc: 0.994 (K=None), 1.000 (DAE), 0.978 (Adaptive+K=S); Meteor Drop: [truncated]iitilde0.057; SBERT ↑: [truncated]iitilde1.227; Ownership R...	Yes; semantic-level embedding; synonym substitution using BERT	Yes; watermark message assigned categorical label (e.g., 4-bit → 1-of-16)	Yes; semantic embeddings via transformer encoder and BERT; SBERT distance as metric

Continued on next page

Table 1 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
Context-aware linguistic steganography model ba... [8]	BERT (encoder), LSTM (decoder)	WMT18 News Commentary (train/test), Yang et al. bits, Doc2Vec, 5,000 stego pairs (8:1:1 split)	BLEU: 30.5, PPL: 22.5, ER: 0.29, KL: 0.02, SIM: 0.86, Stego detection [truncated]iitilde16%	Yes	[Not specified]	GCF (global context), LMR (language model reference), Multi-head attention
DeepTextMark: a deep learning-driven text water... [24]	Model-independent; tested with OPT-2.7B	Dolly ChatGPT (train/validate), C4 (test), robustness & sentence-level test sets	100% accuracy (multi-synonym, 10-sentence), mSMS: 0.9892, TPR: 0.83, FNR: 0.17, Detection: 0.00188s, Insertion: 0.27931s	NO	[Not specified]	[Not specified]
Hi-stega: A hierarchical linguistic steganograp... [37]	GPT-2	Yahoo! News (titles, bodies, comments); 2,400 titles used	ppl: 109.60, MAUVE: 0.2051, ER2: 10.42, Δ(cosine): 0.0088, Δ(simcse): 0.0191	explicit	Social Media	Text
Linguistic steganography: From symbolic space t... [44]	CTRL (generation), BERT (semantic classifier)	5,000 CTRL-generated texts per semanteme (n = 2–16); 1,000 user-generated texts for anti-steganalysis	Classifier Accuracy: 0.9880; Loop Count: 1.0160; PPL: 13.9565; Anti-Steganalysis Accuracy: [truncated]iitilde0.5	implicit	Text	Semanteme (α) as a vector in semantic spac

Continued on next page

Table 1 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
Natural language steganography by chatgpt [34]	[Not specified]	Custom word sets for specific topics (e.g., 16×10-word sets for music reviews)	[Not specified]	Explicit	Specific Genre/Topic Text	Text
Natural language watermarking via paraphraser-b... [28]	Transformer (Paraphraser), BART (BARTScore), BERT (BLEURT, comparisons)	ParaBank2, LS07, CoInCo, Novels, WikiText-2, IMDB, NgNews	LS07 P@1: 58.3, GAP: 65.1; CoInCo P@1: 62.6, GAP: 60.7; Text Recoverability: [truncated]iitilde88–90%	Explicit	[Not specified]	text
Rewriting-Stego: generating natural and control... [18]	BART (bart-base2)	Movie, News, Tweet	BPTS: 4.0, BPTC+S: 4.0, PPL: 62.1, Mean: 44.4, Variance: 2.1e04, Acc: 8.9%	not Explicit	[Not specified]	[Not specified]
ALiSa: Acrostic linguistic steganography based ... [43]	BERT (Google's BERTBase, Uncased)	BookCorpus (10,000 natural texts for evaluation)	PPL: Natural = 13.91, ALiSa = 14.85; LS-RNN/LS-BERT Acc & F1 = [truncated]iitilde0.50; Outperforms GPT-AC/ADG in all cases	No	[Not specified]	[Not specified]

7 RESULTS AND DISCUSSION

This section presents the synthesized findings from the systematic literature review, encompassing 18 primary studies and an additional 14 pending papers. The analysis has been augmented with recent literature from 2024–2025 to address the rapidly evolving nature of this field. The discussion is organized around the six research questions (RQs) and provides a synthesis of trends, quantitative comparisons, and key examples for each. Tables highlight metrics and trade-offs for clarity, with all metrics representing averaged or best-reported values across studies. The analysis contrasts black-box methods (utilizing APIs without internal access) with white-box methods (requiring access to model internals).

7.1 State of Published Literature on LLM-based Steganography (RQ1)

The review identified a significant surge in literature since 2023, with approximately 20 new papers published in 2024–2025 focusing on generative steganography. Early works (pre-2024) primarily concentrated on white-box modifications, such as token sampling in GPT-2, whereas recent trends demonstrate a shift toward hybrid and black-box approaches for more practical, real-world deployment.

Key trends in this evolving field include:

- **Model Preference:** Approximately 70% of studies utilize open-source LLMs such as LLaMA2 and LLaMA3.
- **Overlap with Watermarking:** Approximately 40% of research integrates concepts from digital watermarking.
- **Publication Venues:** Publications are concentrated in preprint servers such as arXiv and conferences including ACL and NeurIPS.

Despite this growth, several gaps persist. Limited focus exists on non-English languages, and only approximately 10% of studies address the ethical implications of these techniques. Recent model examples include **DAIRstega** (2024), which advanced interval-based sampling, and **FreStega** (2024), which provides a plug-and-play approach to imperceptibility.

7.2 Applications of LLM-based Steganographic Techniques (RQ2)

The analysis reveals several distinct applications for LLM-based steganography:

- **Covert Communication:** Approximately 60% of papers focus on this application, particularly for use in censored environments.
- **Watermarking and Fingerprinting:** About 30% of studies use these techniques for content tracing, and 10% focus on fingerprinting LLMs for licensing purposes.

Emerging applications include:

- **Social Media Hiding:** Models such as **Co-Stega** expand text space through context retrieval and entropy enhancement.
- **Jailbreak Attacks:** Steganography can conceal harmful queries, as demonstrated in **StegoAttack**.
- **Data Exfiltration: TrojanStego** embeds secrets directly into LLM outputs.

The field further investigates domain-specific applications, including the utilization of high-entropy texts in news articles and short prompts for question-and-answer paradigms. Additionally, a growing overlap exists with adversarial robustness and potential for multimodal steganography using models such as GPT-4o.

7.3 Evaluation Metrics and Methods for LLM-based Steganography (RQ3)

Performance evaluation for LLM-based steganography relies on three key categories of metrics:

- **Imperceptibility:** Encompasses both **perceptual metrics** (PPL, MAUVE) and **statistical metrics** (KLD, JSD). Cognitive metrics such as BLEU and BERTScore assess semantic similarity.
- **Capacity:** Measured in bits per token/word (bpw/bpt) and embedding rate (ER).
- **Security:** Evaluated through anti-steganalysis accuracy/F1 score and detection rate following attacks.

Evaluation methods encompass automated tools, including steganalysis classifiers, and human fluency judgments. Recent white-box methods such as **ShiMer** achieve a KLD of 0 with a capacity exceeding 2 bpt, whereas black-box methods demonstrate higher PPL (average of 100-300) but provide superior accessibility. For instance, **Ensemble Watermarks** achieves a 98% detection rate but may degrade to 95% following a paraphrase attack. The following table provides a comparison of different methods.

Method Type	**Avg. PPL**	**Avg. KLD**	**Avg. Embed. Rate**	**Human Eval**	**Trend**
Black-box	∼168-363	∼1.76-2.23	∼5.37 bpw	79-91% detection	Higher PPL but robust
White-box	∼3-8	∼0-0.25	∼1.10-5.98 bpt	MAUVE ∼80-92	Lower PPL/KLD, requires internals
Hybrid	N/A	N/A	N/A	95-98% detection post-attack	Balances security but vulnerable

Table 2. Comparison of different LLM-based steganography method types.

A significant need exists for standardized benchmarks, as human evaluations are frequently overlooked in current research.

7.4 Integration of External Knowledge Sources (RQ4)

The integration of external knowledge sources has emerged as a crucial area of research in LLM-based steganography. This integration enhances both capacity and contextual relevance of steganographic systems. Common integrations include:

- **Semantic Resources:** Knowledge graphs and context retrieval, as seen in **Co-Stega**, enhance contextual relevance.
- **Domain Corpora:** Models like **FreStega** use large corpora for distribution alignment.
- **Prompts:** Used to boost entropy and guide text generation.

This integration enhances capacity (e.g., a 15% increase in FreStega) and improves contextual relevance. Although this introduces computational overhead, it remains generally minimal and can be amortized. Future research may explore federated learning to further enhance privacy.

7.5 Limitations and Trade-offs in Current Techniques (RQ5)

Current LLM-based steganographic techniques face several fundamental limitations and trade-offs that constrain their practical deployment and security guarantees:

- **Low Capacity:** Hiding information in short, low-entropy texts (e.g., social media posts) is a significant challenge.
- **Psic Effect:** The Perceptual-Statistical Imperceptibility Conflict Effect (see Section **??**) represents a critical trade-off between perceptual quality and statistical imperceptibility, leading to an average capacity loss of 1–2 bpw when optimizing for PPL over KLD.
- **Vulnerability to Attacks:** Techniques are often vulnerable to paraphrasing and fine-tuning attacks, with detection rates dropping by 5–50% in some cases.
- **Segmentation Ambiguity:** Subword tokenization (e.g., BPE in **SparSamp**) can create ambiguity in message extraction.

- **White-box vs. Black-box Access:** White-box methods offer higher security but require access to model internals, while black-box methods are more practical for real-world deployment but may be less secure.
- **Ethical Concerns:** Issues such as biases, discrimination, and the potential for misuse (e.g., in **TrojanStego**) remain unaddressed in many works.

The following table provides a quantitative overview of these trade-offs.

Limitation/Trade-off	Quantified Impact	Examples
Psic Effect	~1-2 bpw loss	DAIRstega: Higher capacity reduces anti-steg Acc to 58%
Attack Vulnerability	5-50% detection drop	Ensemble WM: 98% to 95%; TrojanStego: 97% to 65%
Entropy/Ambiguity	Capacity cap ~1023 bits	SparSamp: TA reduces accuracy; ShiMer: Cannot boost entropy
Ethical/Overhead	Performance degradation ~5-11%	UTF: HellaSwag drop 5%; FreStega: Needs corpus (100 samples)

Table 3. Key limitations and trade-offs in current LLM-based steganography.

7.6 Future Research Directions (RQ6)

The analysis of current literature and identified limitations reveals several promising avenues for future research in LLM-based steganography:

- **Multimodal Steganography:** Integrating text with other media like images.
- **Robust Defenses:** Developing techniques that are more resilient to attacks, such as paraphrasing.
- **Integration with RAG:** Using Retrieval-Augmented Generation for more adaptive and context-aware systems.
- **Non-English Support:** Expanding research to non-English languages and different cultural contexts.
- **Ethical Frameworks:** Establishing clear guidelines and frameworks to prevent the misuse of these technologies.
- **Provable Security:** Advancing the theoretical foundations to provide stronger security guarantees.
- **Efficient Computation:** Reducing the computational overhead of these techniques.

The field of LLM-based steganography continues to evolve rapidly, with novel models and techniques being developed to address these challenges and explore new possibilities, particularly through the paradigm shift toward context-aware and API-based systems.

8 MAIN FINDINGS

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

8.1 Overview of LLM-based Steganography

The review identifies several important trends in LLM-based linguistic steganography:

- Models like GPT-2, LLaMA, and Baichuan2 serve as foundations for steganographic techniques.
- Both white-box and black-box approaches have emerged with distinct trade-offs.
- Fundamental tensions between imperceptibility, capacity, and security drive ongoing research.

8.2 Key Techniques and Approaches

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

- **LLM-Stega** [38]: Black-box approach using LLM interfaces.
- **Co-Stega**: Context retrieval and entropy enhancement for social media.
- **Zero-shot steganography**: In-context learning with question-answer paradigms.
- **ALiSa**: Token-level embedding in BERT-generated text.

8.3 Critical Challenges

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

- The Psic Effect [42]: A fundamental trade-off between perceptual quality and statistical security (see Section **??**).
- 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.

8.4 Future Outlook

Based on this analysis, several promising directions for future research are identified:

- 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.

9 CONCLUSION

This systematic literature review illuminates the profound impact of Large Language Models (LLMs) on linguistic steganography, demonstrating a clear paradigm shift toward context-aware, generative systems that prioritize imperceptibility, embedding capacity, and naturalness. Through analysis of 18 primary studies (with 14 additional pending for full inclusion), key research questions were addressed, revealing that the published literature is rapidly evolving. Applications now span secure communication in social media, zero-shot generation, and watermarking overlaps.

Evaluation metrics such as Perplexity (PPL), Kullback-Leibler Divergence (KLD), and bits per token/word consistently show LLM-based methods outperforming traditional approaches. This improvement is particularly evident through integration of external semantic resources like context retrieval and domain-specific prompts to enhance relevance and capacity. However, persistent limitations remain, including the Perceptual-Statistical Imperceptibility Conflict (Psic Effect), low entropy in short texts, and challenges in black-box access. These underscore fundamental trade-offs in security and practicality.

The findings establish that contextual compatibility—leveraging domain correlations and communicative patterns—is essential for robust steganographic systems. This development paves the way for more sophisticated covert channels resistant to both human and automated detection. These advancements hold significant implications for information security, enabling high-capacity hidden messaging in everyday digital interactions while mitigating risks such as hallucinations and biases in LLMs.

Future research should concentrate on several key areas: mitigating segmentation ambiguity, developing provably secure black-box frameworks, and exploring multimodal integrations (e.g., text with images) to bridge identified gaps. This review underscores the potential of LLMs to redefine steganography as a cornerstone of secure, imperceptible communication in an increasingly surveilled digital landscape.

REFERENCES

[1] 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[2] 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL] https://arxiv.org/abs/2307.09288

[3] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. ACM, Virtual Event, Canada, 610–623.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023).

[6] Christian Cachin. 1998. An Information-Theoretic Model for Steganography. In *Information Hiding*, David Aucsmith (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 306–318.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] 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.

[9] 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.

[10] 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, San Francisco, CA, USA, 2238–2255.

[11] Jessica Fridrich. 2009. *Steganography in digital media: principles, algorithms, and applications*. Cambridge University Press, Cambridge, UK.

[12] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2019).

[13] Lin Huo and Yu chuan Xiao. 2016. Synonym substitution-based steganographic algorithm with vector distance of two-gram dependency collocations. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. 2776–2780. doi:10.1109/CompComm.2016.7925203

[14] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 2005. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* 62, S1 (08 2005), S63–S63. arXiv:https://pubs.aip.org/asa/jasa/article-pdf/62/S1/S63/11558910/s63_5_online.pdf doi:10.1121/1.2016299

[15] 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*. ACM, 2908–2916.

[16] 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*. ACM, Virtual Event, Republic of Korea, 1529–1548.

[17] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79–86. http://www.jstor.org/stable/2236703

[18] 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.

[19] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, CA, USA, 110–119.

[20] 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*. ACM, Baiona, Spain, 7–12.

[21] Ke Lin, Yiyang Luo, Zijian Zhang, and Ping Luo. 2024. Zero-shot generative linguistic steganography. *arXiv preprint arXiv:2403.10856* (2024).

[22] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (Portland, Oregon) *(HLT '11)*. Association for Computational Linguistics, USA, 142–150.

[23] Mohammed Abdul Majeed, Rossilawati Sulaiman, Zarina Shukur, and Mohammad Kamrul Hasan. 2021. A Review on Text Steganography Techniques. *Mathematics* 9, 21 (2021). doi:10.3390/math9212829

[24] Travis Munyer, Abdullah All Tanvir, Arjon Das, and Xin Zhong. 2024. DeepTextMark: a deep learning-driven text watermarking approach for identifying large language model generated text. *Ieee Access* 12 (2024), 40508–40520.

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, PA, USA, 311–318.

[26] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Chris Callison-Burch, AI Ai2, and Aditya Grover. 2021. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., Virtual Event, 4816–4828.

[27] Yuang 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). Early Access.

[28] 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.

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* (2019). https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[30] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language Models are Unsupervised Multitask Learners*. Technical Report. OpenAI.

[31] De Rosal Ignatius Moses Setiadi, Sudipta Kr Ghosal, and Aditya Kumar Sahu. 2025. AI-Powered Steganography: Advances in Image, Linguistic, and 3D Mesh Data Hiding – A Survey. *Journal of Future Artificial Intelligence and Technologies* 2, 1 (Apr. 2025), 1–23. doi:10.62411/faith.3048-3719-76

[32] Murray Shanahan. 2024. Talking about large language models. *Commun. ACM* 67, 2 (2024), 68–79.

[33] Gustavus J Simmons. 1984. The prisoners' problem and the subliminal channel. In *Advances in Cryptology: Proceedings of Crypto 83*. Springer, Boston, MA, 51–67.

[34] Martin Steinebach. 2024. Natural language steganography by chatgpt. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*. ACM, 1–9.

[35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[37] 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.

[38] 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*. ACM, Melbourne, Australia, 10345–10353.

[39] Jianfei Xiao, Yancan Chen, Yimin Ou, Hanyi Yu, Kai Shu, and Yiyong Xiao. 2024. Baichuan2-Sum: Instruction Finetune Baichuan2-7B Model for Dialogue Summarization. arXiv:2401.15496 [cs.CL] https://arxiv.org/abs/2401.15496

[40] Zhenyu Xu, Ruoyu Xu, and Victor S. Sheng. 2024. Beyond Binary Classification: Customizable Text Watermark on Large Language Models. In *2024 International Joint Conference on Neural Networks (IJCNN)*. 1–8. doi:10.1109/IJCNN60899.2024.10650062

[41] Aiyuan Yang, Bin Xiao, Binyuan Wang, Binxin Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open Large-scale Language Models. *arXiv preprint arXiv:2309.10305* (2023).

[42] 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.

[43] Biao Yi, Hanzhou Wu, Guorui Feng, and Xinpeng Zhang. 2022. ALiSa: Acrostic linguistic steganography based on BERT and Gibbs sampling. *IEEE Signal Processing Letters* 29 (2022), 687–691.

[44] 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.

[45] Si-yu Zhang, Zhongliang Yang, Jinshuai Yang, and Yongfeng Huang. 2021. Provably Secure Generative Linguistic Steganography. *CoRR* abs/2106.02011 (2021). arXiv:2106.02011 https://arxiv.org/abs/2106.02011

[46] Yue Zhang, Siqi Sun, Michel Galley, Chris Brockett, and Jianfeng Gao. 2023. Language Models as Zero-Shot Style Transferers. *arXiv preprint arXiv:2303.03630* (2023).

[47] 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.

[48] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.