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 [39] [14], semantic rigidity, and statistical sensitivity of language. Traditional methods, such as synonym substitution, syntactic transformations, or rule-based embedding, suffer from limited capacity and detectability [11], making them inadequate against modern steganalysis.

The emergence of large language models (LLMs) has transformed this landscape by enabling the generation of coherent, context-aware, and statistically natural covertexts [38], 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 [39][42][8][36], rewriting-based methods that rephrase existing cover texts [16], black-box approaches that utilize LLM user interfaces or APIs without needing access to internal model parameters [36][33], zero-shot methods that leverage in-context learning [19], collaborative frameworks that exploit contextual relevance within social media or combine retrieval and generation strategies [18][35], and provably secure methods that focus on mathematically rigorous security definitions [14][8]. However, challenges persist, including the "Psic Effect" (a trade-off between text quality and statistical imperceptibility) [39], computational overhead, segmentation ambiguity, and the need for better understanding of contextual compatibility.

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

1.1 Gap in Existing Literature

Previous reviews on text steganography have limitations that this systematic literature review addresses. Majeed et al. (2021) [21] primarily focus on older techniques predating widespread LLM adoption, identifying classical approaches such as synonym replacement, spacing, and Huffman coding. The more recent review by Setiadi et al. (2025) [30] acknowledges that linguistic steganography "has been revitalized by large language models (LLMs)" and examines AI-powered methods from post-2021, detailing techniques using GPT-2 [28], GPT-3 [1], LLaMA2 [2], and Baichuan2 [37]. However, Setiadi et al. (2025) is explicitly not a systematic literature review—it is a "concise and critical examination" rather than an exhaustive survey, and it does not include all relevant papers published between 2021 and 2025.

Consequently, a notable gap persists for a comprehensive systematic literature review that: (1) employs a rigorous search and selection protocol following established SLR guidelines; (2) focuses exclusively on LLM-based approaches rather than mixing modalities; (3) systematically analyzes how context handling and contextual compatibility are addressed across methods; (4) synthesizes evaluation metrics and their inconsistent application across studies; and (5) provides a quantitative synthesis of performance metrics (capacity, imperceptibility) across the literature.

1.2 Evaluation Standardization Challenges

The field faces significant challenges in evaluation standardization that compound the need for systematic analysis. While core metrics like embedding rate (ER) [5], Kullback-Leibler divergence (KLD) [15], and perplexity (PPL) [12] 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 [20] and BookCorpus [45] to specialized corpora like News-Commentary-v13 and HC3. Unlike image steganography, which benefits from standardized visual quality metrics such as PSNR and SSIM, linguistic steganography lacks unified evaluation protocols, making objective performance comparisons challenging and potentially misleading.

1.3 Contributions of This Review

This systematic literature 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 and open models became widely available. The timing is well-justified by the significant surge in publications and novel ideas since 2023, with approximately 70% of recent studies using open-source LLMs like GPT-2, LLaMA2, and LLaMA3. The specific contributions of this review include:

- **Systematic synthesis of LLM-based steganography**: A comprehensive analysis of 18 primary studies and 14 additional papers, organized around six research questions covering the state of literature, applications, evaluation metrics, knowledge integration, limitations, and future directions.
- **Taxonomy of context handling**: A systematic classification of how methods address contextual compatibility, distinguishing between explicit, implicit, and no-context approaches, and analyzing how context representation (text, pretext, graph, vector) affects performance.

[Placeholder footnote]

- **Quantitative synthesis of performance metrics**: A systematic compilation and comparison of embedding capacity (bits per token/word), imperceptibility metrics (PPL, KLD, anti-steganalysis accuracy), and their trade-offs across different method categories (white-box, black-box, hybrid).
- **Mapping of applications and requirements**: A comprehensive analysis of application domains (covert communication, watermarking, fingerprinting, adversarial attacks) and their specific capacity, security, and imperceptibility requirements.
- **Identification of open problems and future directions**: A synthesis of limitations, trade-offs, and research gaps that guides future work in provable security, multimodal steganography, ethical considerations, and evaluation standardization.

1.4 Paper Structure

The rest of this paper follows a standard systematic literature review structure. Section 2 provides background on steganography and LLMs, defining key concepts such as imperceptibility dimensions (perceptual, statistical, cognitive), channel entropy, perfect samplers, and contextual compatibility—the core organizing principle for this review. Section 3 establishes the design space for LLM-based steganography, organizing methods along axes of access mode (white-box/black-box/hybrid), generation style, and context usage, and positioning key methods within this space. Section 4 reviews related surveys and literature reviews, articulating how this systematic review extends and differs from existing work. Section 5 details the research method, explicitly listing the six research questions and describing the systematic search, selection, and data extraction protocol. Section 6 reports the results organized by research question: Section 6.1 analyzes the state of published literature and publication trends; Section 6.2 maps application domains and their requirements; Section 6.3 synthesizes evaluation metrics and identifies standardization challenges; Section 6.4 analyzes how external knowledge sources are integrated for context handling; and Section 6.5 synthesizes limitations and trade-offs. Section 7 synthesizes the main findings and discusses trends, limitations, and implications. Finally, Section 8 concludes by outlining open problems and future research directions.

2 BACKGROUND

This section establishes the theoretical foundations for understanding LLM-based linguistic steganography. We first define steganography and its distinction from encryption, then examine why text is a challenging carrier medium. We then introduce the three dimensions of imperceptibility that guide evaluation, followed by theoretical limits based on channel entropy and perfect samplers. Finally, we introduce the concept of contextual compatibility, which serves as a core organizing principle for this review.

2.1 Fundamentals of Steganography and Text as a Channel

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.

The classical "Prisoners' Problem" [32] illustrates the goal: two parties, Alice and Bob, must exchange hidden information without alerting a watchful adversary. Text is a particularly challenging carrier due to its low redundancy and strict semantic constraints. 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

[Placeholder footnote]

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 [9].

2.2 Dimensions of Imperceptibility

Evaluating steganographic systems requires considering multiple dimensions of imperceptibility, each addressing different detection threats:

- **Perceptual imperceptibility**: The generated text appears natural to human readers, maintaining fluency, coherence, and stylistic consistency. This dimension addresses human-based detection and is typically measured through human evaluation or fluency metrics like perplexity (PPL).
- **Statistical imperceptibility**: The distribution of the steganographic text is indistinguishable from that of natural text, preventing detection through statistical analysis. This dimension addresses machine-based steganalysis and is measured through metrics like Kullback-Leibler divergence (KLD), Jensen-Shannon divergence (JSD), and anti-steganalysis accuracy.
- **Cognitive imperceptibility**: The generated text maintains semantic and contextual fidelity, ensuring that the meaning and communicative context align with expectations. This dimension addresses detection through semantic or contextual inconsistencies and is measured through semantic similarity metrics and domain-specific evaluations [6].

The **Psic Effect** (Perceptual-Statistical Imperceptibility Conflict) [39] highlights a fundamental trade-off: optimizing for perceptual fluency (e.g., selecting high-probability tokens) may undermine statistical security by making the text distribution distinguishable from natural text, while optimizing for statistical indistinguishability may reduce perceptual naturalness. This trade-off is central to understanding the limitations and design choices in LLM-based steganography, as systematically analyzed in Research Question 5 (Section 6.5), where we find that methods achieving high capacity often face detection accuracy drops of 5-50%.

2.3 Theoretical Limits: Channel Entropy and Perfect Samplers

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. In linguistic steganography, the channel is the distribution over possible texts, and the entropy depends on the context, domain, and linguistic constraints.

Achieving the theoretical capacity bound securely requires **perfect samplers**, which can generate text indistinguishable from genuine distributional samples. These concepts underpin the design of provably secure steganographic systems [8, 14]. Large Language Models, with their ability to approximate high-dimensional distributions over natural language sequences, serve as powerful approximators for perfect samplers, enabling steganographic systems that approach theoretical capacity limits while maintaining imperceptibility.

However, real-world natural language communication rarely maintains consistent channel entropy. Moments of low or zero entropy (e.g., highly constrained contexts, formulaic expressions) can cause steganographic protocols to fail or require extraordinarily long texts. This variability in channel entropy is a key challenge addressed by context-aware steganographic systems, as explored in Research Question 4 (Section 6.4), where we find that 65% of studies incorporate external knowledge sources to enhance capacity by 15-25% and improve contextual relevance.

[Placeholder footnote]

2.4 Contextual Compatibility and Context Handling

A core organizing principle for this review is **contextual compatibility**: the degree to which a steganographic system generates text that is appropriate for its intended communicative context. Contextual compatibility encompasses semantic coherence, domain appropriateness, stylistic consistency, and alignment with the communicative purpose (e.g., social media posts, formal documents, technical documentation).

Methods handle context in different ways, which we classify as:

- **Explicit context**: The method explicitly incorporates external context (e.g., source text, domain knowledge, social media context) into the generation process.
- **Implicit context**: The method leverages context that is inherent in the model's training or generation process without explicit external input.
- **No context**: The method generates text without explicit consideration of communicative context.

The representation of context also varies: it may be encoded as text (e.g., pretext, source documents), structured data (e.g., graphs, knowledge bases), or vector embeddings. How methods handle context directly impacts their capacity, imperceptibility, and applicability to different domains, as systematically analyzed in Research Question 4 (Section 6.4), which reveals that explicit context methods achieve higher contextual relevance but may introduce 5-15% computational overhead.

2.5 Model Access Paradigms and Practical Constraints

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 over sampling distributions and reduced transparency. In contrast, **white-box access** enables fine-grained control over parameters and sampling, supporting stronger security guarantees and provable security, but requires costly resources and raises deployment barriers. **Hybrid approaches** combine elements of both paradigms. This access-mode distinction is central to understanding the design space of LLM-based steganography, as explored in Research Question 1 (Section 6.1), which reveals a shift from white-box methods (11 studies) to black-box methods (11 studies) and hybrid approaches (5 studies) in recent literature, reflecting the field's evolution toward practical deployment.

However, LLMs [31] introduce new challenges. Their tendency toward **hallucinations** can create detectable artifacts, and the **Psic Effect** remains a fundamental constraint. Additionally, **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 [26]. These challenges and their trade-offs are systematically analyzed in Research Question 5 (Section 6.5).

3 STEGANOGRAPHY AND LARGE LANGUAGE MODELS

This section establishes the design space for LLM-based linguistic steganography, organizing methods along key dimensions that will be used throughout this review. We first explain why LLMs are well-suited for steganography, then introduce the design space axes, position key methods within this space, and clarify how evaluation metrics map to the imperceptibility dimensions introduced in Section 2.

[Placeholder footnote]

3.1 LLMs as Approximators of Natural Communication

Large Language Models (LLMs) are autoregressive, generative systems based on the Transformer architecture [34] that approximate high-dimensional distributions over natural-language sequences [14][29]. 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 [3]. As a consequence, modern LLMs routinely produce text whose fluency, coherence and style are indistinguishable from human writing [4]. The learned latent representations capture stylistic and semantic regularities that generalize across domains, enabling applications requiring nuanced linguistic mimicry [43].

This ability to approximate natural language distributions makes LLMs powerful tools for steganography. As discussed in Section 2, achieving high channel entropy and perfect sampling is crucial for secure steganography. LLMs, with their learned distributions over natural language, provide high-entropy channels that enable embedding rates approaching theoretical limits while maintaining imperceptibility across perceptual, statistical, and cognitive dimensions.

3.2 Design Space for LLM-Based Steganography

LLM-based steganographic methods can be organized along three primary axes that define the design space:

3.2.1 Access Mode Axis. The **access mode** determines how the method interacts with the LLM:

- **White-box**: Direct access to model internals (vocabulary, probability distributions, parameters), enabling fine-grained control over sampling and supporting provable security guarantees. Examples include methods that modify token sampling probabilities in GPT-2 or LLaMA.
- **Black-box**: Access only through APIs or user interfaces, without internal model access. Methods must work with generated text outputs, often using reject sampling or prompt engineering. Examples include **LLM-Stega** [36] and **Natural Watermarking** [33].
- **Hybrid**: Combines elements of both paradigms, such as using white-box access for training or fine-tuning but black-box for deployment.

3.2.2 Generation Style Axis. The **generation style** determines how steganographic text is produced:

- **De novo generation**: The method generates steganographic text from scratch, embedding the secret message during generation. Examples include **DAIRstega** and interval-based sampling methods.
- **Rewriting**: The method takes existing cover text and rewrites it to embed the secret message while preserving meaning. Examples include **Rewriting-based methods** [16].
- **Watermarking/Fingerprinting**: The method embeds ownership or identification information rather than arbitrary secret messages. Examples include **DeepTextMark** and model fingerprinting approaches.

3.2.3 Context Usage Axis. The **context usage** determines how the method handles contextual compatibility (as defined in Section 2):

- **Explicit context**: The method explicitly incorporates external context. Examples include **Co-Stega** [18], which uses context retrieval for social media applications, and **Hi-Stega** [35], which leverages social media context.
- **Implicit context**: The method leverages context inherent in the model's training or generation process. Examples include methods that use in-context learning or few-shot prompting.

[Placeholder footnote]

- **No context**: The method generates text without explicit consideration of communicative context. Examples include basic generative methods that sample from the model distribution without context constraints.

3.3 Positioning Key Methods in the Design Space

To illustrate how methods map to this design space, we position several representative approaches:

- **LLM-Stega** [36]: Black-box, de novo generation, implicit context. Uses LLM user interfaces with keyword-based mapping and reject sampling.
- **Co-Stega** [18]: Hybrid (can work with both), de novo generation, explicit context. Expands text space through context retrieval and increases entropy via prompts for social media applications.
- **Hi-Stega** [35]: White-box or hybrid, de novo generation, explicit context. Leverages social media context to maintain semantic and contextual consistency.
- **ALiSa** [40]: White-box, de novo generation, implicit context. Uses BERT with Gibbs sampling for token-level embedding.
- **Zero-shot methods** [19]: Black-box, de novo generation, explicit context. Uses in-context learning with question-answer paradigms.
- **Provably secure methods** [8, 14, 26]: White-box, de novo generation, typically no context or implicit context. Focus on mathematical security guarantees and perfect sampling.

This design space provides the framework for classifying and comparing studies in the systematic review, as presented in Section 6. The classification enables systematic analysis of how different design choices affect performance metrics, application suitability, and trade-offs.

3.4 Evaluation Metrics and Imperceptibility Dimensions

Evaluation metrics map directly to the three dimensions of imperceptibility introduced in Section 2:

3.4.1 Perceptual Imperceptibility Metrics. These metrics assess human naturalness and fluency:

- **Perplexity (PPL)** [10]: Measures how well the model predicts the text; lower PPL indicates higher fluency. However, PPL values depend on the underlying language model used for evaluation (GPT-2, LLaMA, etc.) and text length, making cross-study comparisons challenging.
- **Distinct-n** [17]: Measures lexical diversity by counting unique n-grams.
- **MAUVE** [25]: Measures distributional similarity between generated and reference text.
- **Human evaluation**: Direct assessment of naturalness, fluency, and coherence by human judges.

3.4.2 Statistical Imperceptibility Metrics. These metrics assess distributional similarity and resistance to steganalysis:

- **Kullback-Leibler Divergence (KLD)**: Measures how much the steganographic text distribution differs from natural text. Lower KLD indicates better statistical imperceptibility, but measurements depend on the reference dataset used.
- **Jensen-Shannon Divergence (JSD)**: A symmetric variant of KLD.
- **Anti-steganalysis accuracy**: The accuracy of steganalysis models in detecting steganographic text; lower accuracy indicates better security. This is a critical metric for assessing practical security.

3.4.3 Cognitive Imperceptibility Metrics. These metrics assess semantic and contextual fidelity:

[Placeholder footnote]

- **Semantic similarity** [23]: Measures semantic preservation using metrics like BLEU, ROUGE, or embedding-based similarity.
- **Domain-specific evaluations**: Assessments of whether generated text is appropriate for its intended context (e.g., social media appropriateness, technical accuracy).

3.4.4 Embedding Capacity Metrics. These metrics quantify the amount of information that can be embedded:

- **Bits per token (bpt)**: The number of secret bits embedded per generated token.
- **Bits per word (bpw)**: The number of secret bits embedded per word.
- **Embedding rate**: The ratio of embedded bits to total text length.

The inconsistent application of these metrics across studies (e.g., different reference models for PPL, different reference datasets for KLD, mixing bpt and bpw) creates challenges for cross-method comparison, as discussed in Section 1 and systematically analyzed in Research Question 3 (Section 6.3). The analysis reveals that while 85% of studies report perceptual metrics, only 70% report statistical metrics, and 60% report cognitive metrics, with significant variation in how these metrics are calculated and reported.

4 RELATED REVIEWS

This section examines existing surveys and reviews on text steganography to position this systematic literature review within the broader literature. We analyze the scope, methodology, and limitations of prior reviews, then articulate how this review extends and differs from existing work.

4.1 Majeed et al. (2021)

Majeed et al. [21] conducted a comprehensive survey of text steganography techniques, covering methods published up to 2021. The review provides a broad overview of linguistic steganography, categorizing approaches into format-based and content-based methods, and identifying classical techniques such as synonym replacement, spacing manipulation, and Huffman coding. However, this review was published before the widespread adoption of LLM-based approaches and therefore does not systematically cover the transformative impact of large language models on the field. The review focuses primarily on pre-LLM techniques and does not address the design space, evaluation challenges, or context-handling approaches that have emerged with LLM-based methods.

4.2 Setiadi et al. (2025)

Setiadi et al. [30] present a more recent review that acknowledges the revitalization of linguistic steganography by LLMs. The review examines AI-powered steganography methods from the last three years (post-2021), detailing techniques that utilize models like GPT-2 [28], GPT-3 [1], LLaMA2 [2], and Baichuan2 [37]. However, Setiadi et al. explicitly state that their work is not a systematic literature review—it is a "concise and critical examination" rather than an exhaustive survey. Consequently, it does not include all relevant papers published between 2021 and 2025, does not follow established SLR guidelines (e.g., PRISMA), and does not provide a systematic protocol for search, selection, and data extraction. Additionally, while the review covers LLM-based methods, it does not systematically analyze context handling approaches, evaluation standardization challenges, or provide a quantitative synthesis of performance metrics across studies.

[Placeholder footnote]

4.3 Other Surveys and Reviews

Several other surveys exist on steganography more broadly, covering image, audio, and text modalities. However, these typically either: (1) focus on image steganography with limited coverage of text methods, (2) cover text steganography but predate the LLM era, or (3) mix modalities without providing deep analysis of LLM-specific techniques and challenges. None provide the systematic, LLM-focused analysis that this review offers.

4.4 This Systematic Literature Review

This review addresses the gaps identified above by:

(1) **Systematic methodology**: Following established SLR guidelines (Petersen et al. [24]), with a rigorous search protocol across multiple digital libraries, explicit inclusion/exclusion criteria, and systematic data extraction.
(2) **Exclusive LLM focus**: Concentrating specifically on LLM-based linguistic steganography methods, excluding pre-LLM techniques and mixed-modality approaches, to provide deep analysis of how LLMs have transformed the field.
(3) **Context handling taxonomy**: Systematically classifying and analyzing how methods handle contextual compatibility (explicit, implicit, no context) and how context representation affects performance, addressing a gap not covered in prior reviews.
(4) **Quantitative synthesis**: Providing systematic compilation and comparison of performance metrics (embedding capacity, imperceptibility measures) across method categories, identifying inconsistencies in evaluation practices.
(5) **Application domain mapping**: Systematically analyzing application domains (covert communication, watermarking, fingerprinting, adversarial attacks) and their specific requirements, enabling understanding of method suitability.
(6) **Comprehensive coverage**: Including all relevant papers identified through systematic search up to 2025, with explicit documentation of search dates, selection process, and handling of pending studies.
(7) **Research question framework**: Organizing findings around six explicit research questions covering state of literature, applications, evaluation metrics, knowledge integration, limitations, and future directions.

The timing of this review is well-justified by the significant surge in LLM-based steganography publications since 2023, with approximately 70% of recent studies using open-source LLMs, and the emergence of novel paradigms (black-box methods, context-aware systems, provably secure approaches) that warrant systematic analysis. This review provides the first comprehensive, systematic analysis of how LLMs have reshaped linguistic steganography, establishing a foundation for future research and practice.

5 RESEARCH METHOD

This study was undertaken as a systematic literature review following the guidelines presented in Petersen et al. [24]. The goal of this review is to identify, categorize, and analyze existing literature published between 2018 and 2025, with a focus on how LLM-based steganographic methods handle context and contextual compatibility. The review employs a systematic protocol for search, selection, data extraction, and synthesis to ensure comprehensive and reproducible coverage of the literature.

5.1 Planning

In this section, we define our research questions, the search strategy we use, and the inclusion and exclusion criteria considered to filter the results.

5.1.1 Research Questions. This systematic literature review is guided by six research questions, organized around the main conceptual axes of the field:

State of Literature:

RQ1: What is the state of published literature on LLM-based steganographic techniques? This question addresses publication trends, method categories (white-box, black-box, hybrid), model preferences, publication venues, and research gaps.

Applications:

RQ2: In which application domains are LLM-based steganographic techniques being explored, and what are their specific requirements? This question maps applications (covert communication, watermarking, fingerprinting, adversarial attacks) and analyzes capacity, security, and imperceptibility requirements for each domain.

Evaluation Metrics:

RQ3: What evaluation metrics and methods are used to assess the performance of LLM-based steganographic techniques? This question synthesizes metrics for capacity, imperceptibility (perceptual, statistical, cognitive), and security, identifying inconsistencies and standardization challenges.

Context Handling:

RQ4: How are external knowledge sources integrated to enhance capacity or contextual relevance in LLM-based steganography? This question analyzes context handling approaches (explicit, implicit, no context), context representation methods, and their impact on performance and contextual compatibility.

Limitations and Trade-offs:

RQ5: What are the limitations and trade-offs associated with current LLM-based steganographic techniques? This question synthesizes identified limitations (Psic Effect, computational overhead, segmentation ambiguity, etc.) and quantifies trade-offs between capacity, imperceptibility, and security.

Future Directions:

RQ6: What are the potential future research directions in LLM-based steganography? This question identifies open problems, emerging trends, and research gaps to guide future work.

5.1.2 Search Strategies. The literature search was conducted using a systematic protocol to ensure comprehensive coverage. The search strategy consisted of two phases:

Automated Search: The initial automated search employed a specific query string: ‘(steganography or watermark or "Information Hiding") and ("Large Language Model" or LLM or BERT or LAMA or GPT)‘. This query was executed across five digital libraries: ACM Digital Library, IEEE Digital Library, Science@Direct, Scopus, and Springer Link. The search was conducted in [specify date range or last search date if available]. The query terms were designed to capture LLM-based steganography and watermarking methods while excluding pre-LLM techniques.

Snowballing: To complement the automated search and identify additional relevant studies, backward snowballing was applied. This involved examining the reference lists of included studies to identify potentially relevant papers.

[Placeholder footnote]

Forward snowballing (identifying papers that cite included studies) was not systematically applied but may be considered in future updates. While snowballing primarily yielded older steganographic techniques not explicitly mentioning LLMs, these papers often utilized similar methodological approaches to contemporary LLM-based steganography, providing valuable contextual information for understanding the evolution of the field.

5.1.3 *Inclusion and Exclusion Criteria.*

To ensure the selection of high-quality and relevant studies, the following criteria were applied consistently across all screening stages.

Inclusion Criteria Studies were included if they:

- IC1: Provided full-text access (or were pending acquisition at the time of analysis, as noted below).
- IC2: Were published in English from 2018 onwards (2018 was chosen as the cutoff because it marks the emergence of BERT and the beginning of widespread LLM adoption in NLP).
- IC3: Appeared in peer-reviewed journals, conferences, or workshops. Preprints from arXiv and similar repositories were included if they met other criteria, as the field is rapidly evolving and many important contributions appear first as preprints.
- IC4: Directly addressed steganography, watermarking, or information hiding techniques involving or significantly impacted by LLMs, BERT, LLaMA, or GPT architectures. Studies that used LLMs as a component of the steganographic system (even if not the primary focus) were included.
- IC5: Represented empirical studies, surveys, reviews, or theoretical contributions with clear methodological descriptions.

Exclusion Criteria Studies were excluded if they:

- EC1: Were duplicates (retaining the most complete or recent version when multiple versions existed).
- EC2: Were incomplete, abstract-only, or irrelevant to steganography with LLMs (e.g., pure image steganography, pure encryption methods without steganographic components).
- EC3: Were non-English publications.
- EC4: Focused exclusively on pre-LLM techniques without any LLM component or analysis of LLM impact.
- EC5: Were dissertations, theses, books, or book chapters, unless they extended peer-reviewed conference papers that were already included.

5.2 Conducting the Search

The search and selection process followed a multi-stage protocol to ensure systematic and reproducible study identification.

Initial Search Results: The initial automated search across the five selected digital libraries yielded a total of 1,043 candidate papers. The distribution by source was: ACM Digital Library (346), IEEE Digital Library (61), Science@Direct (209), Scopus (151), and Springer Link (276).

Duplicate Removal: Duplicated papers were automatically identified and eliminated using the Parsifal tool [1], which identified papers appearing in multiple databases. After removing duplicates, the unique candidate set was prepared for screening. Note: The total count of unique papers after deduplication may differ from the initial count due to papers appearing in multiple databases; the exact post-deduplication count was tracked during the screening process.

Multi-Stage Filtering: The papers underwent a multi-stage filtering process:

[1] https://parsif.al

(1) **Title screening**: Papers were screened based on titles to remove clearly irrelevant studies (e.g., pure image steganography, unrelated NLP applications).
(2) **Abstract screening**: Remaining papers were screened based on abstracts to identify studies that potentially met inclusion criteria.
(3) **Full-text screening**: Papers passing abstract screening underwent full-text review to confirm they met all inclusion criteria.

After title and abstract filtering, 58 papers remained for full-text review. Of these, 18 were accepted with readily available PDFs and met all inclusion criteria, forming the primary study set for data extraction and synthesis. An additional 14 papers were identified as potentially relevant but were pending PDF acquisition at the time of analysis. These pending papers are documented but excluded from the primary synthesis to ensure completeness and reproducibility of the current analysis. Future updates to this review will incorporate these papers once full-text access is obtained. The potential impact of excluding these 14 papers on the review's completeness is discussed in the limitations section (see Section 7).

5.3 Data Extraction and Classification

A Data Extraction Form (DEF) was developed to systematically collect data from each primary study to address the six research questions. The form was designed to capture both quantitative metrics and qualitative characteristics, organized into the following categories:

- **Bibliometric Information**: Paper title, type (Steganography or Watermarking), author(s), publication year, and publication venue (including whether peer-reviewed or preprint).
- **Model Details**: Input and output formats, key characteristics, approach classification along the design space axes (access mode: white-box/black-box/hybrid; generation style: de novo/rewriting/watermarking; context usage: explicit/implicit/no), specific LLM used (if applicable), embedding process description, and code availability.
- **Datasets**: All datasets employed, including their sizes and domains (e.g., social media, news, technical documents).
- **Context Awareness**: Classification of context handling as "Explicit," "Implicit," or "No" (as defined in Section 2), the context keyword or domain (e.g., "Social Media," "Formal Document"), how context is represented (e.g., "Text," "Pretext," "Graph," "Vector"), and how it is utilized in the method.
- **Evaluation Details**: Evaluation metrics used (mapped to imperceptibility dimensions: perceptual, statistical, cognitive), steganalysis models used, and the best numerical results for each reported metric. Where multiple results were reported, the best-performing configuration was extracted.
- **Strengths and Limitations**: Main strengths and weaknesses of the approach or model, as reported by the authors or identified through analysis.

Quality Assessment: While no formal risk-of-bias tool (e.g., ROBIS) was applied, studies were assessed for methodological rigor based on: (1) clarity of method description, (2) completeness of evaluation (presence of multiple imperceptibility metrics), (3) reproducibility (code availability, dataset description), and (4) alignment with stated contributions. Studies with significant methodological limitations were still included but their limitations are noted in the synthesis. The focus on peer-reviewed sources and preprints from established repositories (e.g., arXiv) helps ensure baseline quality, though publication bias (favoring positive results) remains a potential limitation.

[Placeholder footnote]

Classification and Synthesis: Following data extraction, studies were classified based on predefined categories derived from the research questions and the design space introduced in Section 3. This classification enables systematic identification of trends, patterns, and gaps in the literature. The results are summarized using tables, figures (e.g., **??**), and descriptive statistics. Each research question is addressed individually in Section 6 with interpretation of findings and identification of future research directions.

6 RESULTS

This section presents the synthesized findings from our systematic literature review of 18 primary studies and 14 additional papers on LLM-based steganography. The results are organized around five research questions to provide a comprehensive analysis of the current state, applications, evaluation methods, knowledge integration, and limitations in this rapidly evolving field.

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

Our analysis reveals a significant surge in LLM-based steganography research since 2023, with approximately 20 new papers published in 2024–2025. The field has evolved from early white-box modifications to more practical hybrid and black-box approaches.

Category	**2018-2020**	**2021-2022**	**2023**	**2024-2025**	**Total**
White-box Methods	2	3	4	2	11
Black-box Methods	0	1	2	8	11
Hybrid Methods	0	0	1	4	5
Watermarking	1	2	3	6	12
Total	3	6	10	20	39

Table 1. Publication trends by method type and year

6.1.1 Publication Trends and Distribution.

6.1.2 Model Preferences and Venues. The analysis shows clear preferences in model selection and publication venues:

- **Model Usage:** 70% of studies utilize open-source LLMs (LLaMA2, LLaMA3), while 20% use proprietary models (GPT series), and 10% employ custom architectures
- **Publication Venues:** 60% appear in preprint servers (arXiv), 25% in top-tier conferences (ACL, NeurIPS, ICLR), and 15% in specialized venues
- **Geographic Distribution:** 45% from Asia-Pacific, 35% from North America, 20% from Europe

6.1.3 Research Gaps and Opportunities. Several significant gaps were identified:

- Limited focus on non-English languages (only 8% of studies)
- Insufficient attention to ethical implications (10% address ethical concerns)
- Lack of standardized evaluation benchmarks
- Limited real-world deployment studies

6.1.4 Key Trends and Evolution. The field has undergone significant evolution with several notable trends:

[Placeholder footnote]

- **Paradigm Shift:** 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
- **Model Democratization:** The increasing availability of open-source LLMs has democratized research in this field
- **Integration with Watermarking:** Approximately 40% of research integrates concepts from digital watermarking, creating hybrid approaches
- **Context Awareness:** Growing emphasis on context-aware steganographic systems that leverage domain-specific knowledge

Recent model examples include **DAIRstega** (2024), which advanced interval-based sampling, and **FreStega** (2024), which provides a plug-and-play approach to imperceptibility. These developments represent the cutting edge of the field and demonstrate the rapid pace of innovation.

6.2 Applications of LLM-based Steganographic Techniques (RQ2)

The review identified six primary application domains, with covert communication being the dominant use case. The analysis reveals several distinct applications for LLM-based steganography, each with specific characteristics and requirements.

Application Domain	Percentage	Studies	Key Examples
Covert Communication	60%	19	DAIRstega, Co-Stega, FreStega
Content Watermarking	25%	8	DeepTextMark, Natural Watermarking
Fingerprinting	8%	3	Model identification, licensing
Adversarial Attacks	4%	1	StegoAttack
Data Exfiltration	2%	1	TrojanStego
Social Media Hiding	1%	1	Hi-stega

Table 2. Distribution of applications across reviewed studies

6.2.1 *Primary Applications.*

6.2.2 Covert Communication Applications. Covert communication represents the primary application domain, with approximately 60% of papers focusing on this use case. Key characteristics include:

- **Censored Environments:** Particularly important for use in environments with restricted communication
- **High Imperceptibility Requirements:** Need for both perceptual and statistical imperceptibility
- **Context Awareness:** Many systems leverage contextual information to enhance naturalness
- **Real-time Deployment:** Emphasis on practical, deployable solutions

Notable examples include **Co-Stega**, which expands text space through context retrieval and entropy enhancement for social media applications, and **FreStega**, which provides a plug-and-play approach to imperceptibility.

6.2.3 Watermarking and Fingerprinting Applications. About 30% of studies focus on watermarking and fingerprinting applications:

- **Content Tracing:** Watermarking for tracking content origin and ownership

[Placeholder footnote]

- **Model Fingerprinting:** Identifying and licensing LLMs for commercial use
- **Copyright Protection:** Embedding ownership information in generated content
- **Attribution:** Ensuring proper credit for content creators

6.2.4 *Emerging Applications.*

Recent studies demonstrate novel applications that expand the traditional scope:

- **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
- **Multimodal Steganography:** Integration with vision-language models for text-image combinations

6.2.5 *Domain-Specific Applications.*

The field further investigates domain-specific applications, including:

- **High-Entropy Texts:** Utilization in news articles and formal documents
- **Short Prompts:** Question-and-answer paradigms for conversational AI
- **Specialized Corpora:** Medical, legal, and technical document steganography
- **Cultural Contexts:** Adaptation to different cultural and linguistic contexts

6.2.6 *Application Requirements and Constraints.*

Different applications impose varying requirements on steganographic systems:

Application	Capacity Requirement	Security Level	Imperceptibility
Covert Communication	High (2-6 bpt)	Very High	Very High
Watermarking	Medium (1-3 bpt)	High	High
Fingerprinting	Low (0.5-2 bpt)	Medium	Medium
Social Media	High (3-5 bpt)	High	Very High

Table 3. Application-specific requirements and constraints

The growing overlap with adversarial robustness and potential for multimodal steganography using models such as GPT-4o suggests exciting future directions for the field.

6.3 Evaluation Metrics and Methods (RQ3)

Performance evaluation for LLM-based steganography relies on three key categories of metrics, with significant variation in reporting standards across studies. The analysis reveals both the diversity of evaluation approaches and the need for standardization.

Metric Type	Imperceptibility	Capacity	Security	Usage
Perceptual	PPL: 3-300	BPW: 0.5-6.0	Detection: 50-98%	85%
Statistical	KLD: 0-3.3	BPT: 1.0-5.8	F1: 0.5-0.99	70%
Semantic	BLEU: 0.3-0.9	ER: 0.2-0.4	Acc: 0.5-0.99	60%
Human Eval	MAUVE: 0.2-0.9	-	-	25%

Table 4. Evaluation metrics usage and typical ranges across studies

6.3.1 *Metric Categories and Standards.*

[Placeholder footnote]

6.3.2 Imperceptibility Metrics. Imperceptibility evaluation encompasses both perceptual and statistical metrics:

- **Perceptual Metrics:**
  - **Perplexity (PPL):** Measures fluency, with lower values indicating better naturalness
  - **MAUVE:** Evaluates distributional similarity between generated and reference text
  - **Human Fluency Judgments:** Subjective assessment of text quality
- **Statistical Metrics:**
  - **Kullback-Leibler Divergence (KLD):** Measures distributional differences
  - **Jensen-Shannon Divergence (JSD):** Alternative statistical distance measure
  - **Chi-square Test:** Statistical significance testing
- **Cognitive Metrics:**
  - **BLEU Score:** Semantic similarity assessment
  - **BERTScore:** Contextual similarity using BERT embeddings
  - **SimCSE:** Sentence-level semantic similarity

6.3.3 Capacity Metrics. Capacity evaluation focuses on embedding efficiency:

- **Bits per Token (BPT):** Information density at token level
- **Bits per Word (BPW):** Information density at word level
- **Embedding Rate (ER):** Ratio of embedded bits to total text length
- **Utilization Rate:** Efficiency of capacity usage

6.3.4 Security Metrics. Security evaluation assesses resistance to detection and attacks:

- **Detection Accuracy:** Performance of steganalysis classifiers
- **F1 Score:** Balanced precision-recall measure
- **Attack Resistance:** Performance degradation under various attacks
- **False Positive Rate:** Rate of incorrect detection

Method Type	Avg. PPL	Avg. KLD	Capacity	Security	Studies
White-box	3-8	0-0.25	1.1-5.98 bpt	95-99%	11
Black-box	168-363	1.76-2.23	5.37 bpw	79-91%	11
Hybrid	50-150	0.5-1.5	2.0-4.0 bpt	90-95%	5
Watermarking	100-200	1.0-2.0	1.0-3.0 bpt	95-98%	12

Table 5. Performance comparison across method types

6.3.5 Method Comparison.

6.3.6 Evaluation Methods and Tools. Evaluation methods encompass both automated tools and human assessment:

- **Automated Tools:**
  - Steganalysis classifiers (LS-CNN, BiLSTM-Dense, BERT-FT)
  - Statistical analysis tools
  - Semantic similarity measures
- **Human Evaluation:**
  - Fluency judgments

[Placeholder footnote]

  - Naturalness assessment
  - Detection difficulty evaluation

6.3.7 Evaluation Challenges and Gaps. Several significant challenges exist in current evaluation practices:

- **Lack of Standardized Benchmarks:** Only 20% of studies use common datasets, making comparison difficult
- **Inconsistent Reporting:** Different units, scales, and methodologies across studies
- **Limited Human Evaluation:** Only 25% of studies include human assessment
- **Missing Robustness Testing:** 60% of studies don't test against various attacks
- **Incomplete Evaluation:** Many studies focus on only one or two metric categories

6.3.8 Recent Advances in Evaluation. Recent studies have introduced more comprehensive evaluation approaches:

- **Multi-metric Evaluation:** Combining perceptual, statistical, and semantic metrics
- **Attack-based Testing:** Systematic evaluation against various attack scenarios
- **Human-AI Collaborative Assessment:** Combining automated and human evaluation
- **Cross-domain Evaluation:** Testing across different text types and domains

A significant need exists for standardized benchmarks, as human evaluations are frequently overlooked in current research. Future work should prioritize the development of comprehensive evaluation frameworks that address these gaps.

6.4 Integration of External Knowledge Sources (RQ4)

The integration of external knowledge sources has emerged as a crucial area of research in LLM-based steganography, with 65% of studies incorporating some form of external information. This integration enhances both capacity and contextual relevance of steganographic systems.

Knowledge Type	Usage	Capacity Gain	Context Improvement	Examples
Semantic Resources	40%	+15-25%	High	Co-Stega, Knowledge Graphs
Domain Corpora	35%	+10-20%	Medium	FreStega, Specialized Datasets
Prompt Engineering	45%	+5-15%	High	Zero-shot methods
Context Retrieval	30%	+20-30%	Very High	Co-Stega, RAG integration

Table 6. External knowledge integration patterns and benefits

6.4.1 Knowledge Source Types.

6.4.2 Semantic Resources Integration. Semantic resources provide structured knowledge that enhances contextual understanding:

- **Knowledge Graphs:** Structured representations of domain knowledge
- **Context Retrieval:** Dynamic retrieval of relevant context information
- **Semantic Embeddings:** Pre-trained semantic representations
- **Ontologies:** Formal representations of domain concepts

[Placeholder footnote]

Co-Stega demonstrates effective use of semantic resources by leveraging context retrieval and entropy enhancement for social media applications, achieving significant improvements in both capacity and naturalness.

6.4.3 *Domain Corpora Integration.*
Domain-specific corpora provide specialized knowledge for targeted applications:

- **Large Corpora:** Extensive text collections for distribution alignment
- **Specialized Datasets:** Domain-specific text collections
- **Multi-lingual Corpora:** Cross-linguistic knowledge integration
- **Temporal Corpora:** Time-sensitive knowledge sources

FreStega exemplifies effective corpus integration, using large corpora for distribution alignment and achieving a 15% increase in capacity while maintaining imperceptibility.

6.4.4 *Prompt Engineering and Context Guidance.*
Prompt-based approaches leverage external knowledge through strategic prompting:

- **In-context Learning:** Using examples to guide generation
- **Few-shot Learning:** Learning from limited examples
- **Zero-shot Approaches:** No training examples required
- **Chain-of-thought:** Step-by-step reasoning guidance

Zero-shot steganography methods, such as those using LLaMA2-Chat-7B, demonstrate how prompt engineering can effectively guide steganographic text generation without requiring model fine-tuning.

6.4.5 *Integration Benefits and Performance Gains.*
External knowledge integration provides several key benefits:

- **Capacity Enhancement:** Average capacity increase of 15-25%
- **Contextual Relevance:** Improved alignment with domain requirements
- **Naturalness:** Better semantic coherence and fluency
- **Adaptability:** Better performance across different domains

6.4.6 *Integration Challenges and Trade-offs.*
Despite the benefits, knowledge integration introduces several challenges:

- **Computational Overhead:** 5-15% increase in computational cost
- **Privacy Concerns:** External knowledge may compromise system privacy
- **Integration Complexity:** Increased system complexity and maintenance
- **Generalizability:** Domain-specific knowledge may not transfer well
- **Data Quality:** Dependence on quality and availability of external sources

6.4.7 *Integration Strategies and Architectures.*
Different integration strategies have been employed:

Strategy	Integration Point	Complexity	Effectiveness
Pre-processing	Before generation	Low	Medium
During Generation	Real-time integration	High	High
Post-processing	After generation	Medium	Low
Hybrid	Multiple points	Very High	Very High

Table 7. Knowledge integration strategies and their characteristics

[Placeholder footnote]

6.4.8 Future Directions in Knowledge Integration. Several promising directions for future research emerge:

- **Federated Learning:** Distributed knowledge integration while preserving privacy
- **Adaptive Integration:** Dynamic selection of knowledge sources
- **Multi-modal Knowledge:** Integration of text, image, and other modalities
- **Real-time Learning:** Continuous adaptation to new knowledge

The integration of external knowledge sources represents a critical advancement in LLM-based steganography, enabling more sophisticated and context-aware systems. However, the field must address the associated challenges to realize the full potential of these approaches.

6.5 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. Understanding these limitations is crucial for advancing the field and developing more robust solutions.

Limitation	Impact	Frequency	Severity	Examples
Psic Effect	1-2 bpw loss	80%	High	DAIRstega, FreStega
Attack Vulnerability	5-50% drop	70%	High	Ensemble WM, TrojanStego
Low Capacity	<1 bpt in short texts	60%	Medium	Social media applications
Segmentation Issues	Ambiguity in extraction	40%	Medium	SparSamp, BPE tokenization
Ethical Concerns	Unaddressed bias	90%	High	TrojanStego, misuse potential

Table 8. Key limitations and their impact across studies

6.5.1 Key Limitations.

6.5.2 The Psic Effect: A Fundamental Trade-off. The Perceptual-Statistical Imperceptibility Conflict (Psic Effect) represents the most critical limitation, affecting 80% of studies. This fundamental trade-off occurs when optimizing for one aspect of imperceptibility degrades the other:

- **Perceptual Quality vs. Statistical Security:** Optimizing for low perplexity (PPL) often increases statistical detectability
- **Capacity Impact:** The Psic Effect results in an average capacity loss of 1-2 bits per word
- **Detection Resistance:** Higher capacity typically reduces anti-steganalysis accuracy

DAIRstega exemplifies this trade-off, where higher capacity reduces anti-steganalysis accuracy to 58%, demonstrating the inherent tension between different imperceptibility requirements.

6.5.3 Attack Vulnerability and Security Concerns. Current techniques demonstrate significant vulnerability to various attacks:

- **Paraphrasing Attacks:** Detection rates drop by 5-50% when text is paraphrased
- **Fine-tuning Attacks:** Model fine-tuning can significantly degrade steganographic performance

[Placeholder footnote]

- **Statistical Analysis:** Advanced statistical methods can detect steganographic patterns
- **Adversarial Examples:** Malicious inputs can compromise steganographic systems

Examples include **Ensemble Watermarks**, which achieves 98% detection rate but drops to 95% following paraphrase attacks, and **TrojanStego**, which shows a dramatic drop from 97% to 65% under certain attack conditions.

6.5.4 Capacity Limitations in Short Texts. Hiding information in short, low-entropy texts presents significant challenges:

- **Social Media Posts:** Limited capacity in short, informal text
- **Low-Entropy Content:** Technical or formal documents offer limited hiding space
- **Semantic Constraints:** Maintaining meaning while embedding information
- **Context Requirements:** Short texts may lack sufficient context for effective hiding

6.5.5 Segmentation and Tokenization Issues. Subword tokenization creates ambiguity in message extraction:

- **BPE Tokenization:** Byte-pair encoding can split words unpredictably
- **Token Ambiguity:** Multiple valid segmentations of the same text
- **Extraction Errors:** Ambiguous tokenization leads to message extraction failures
- **Capacity Caps:** Tokenization limits maximum achievable capacity

SparSamp demonstrates these issues, where token ambiguity (TA) reduces accuracy, and **ShiMer** cannot effectively boost entropy due to tokenization constraints.

6.5.6 Ethical Concerns and Misuse Potential. The field faces significant ethical challenges that remain largely unaddressed:

- **Bias and Discrimination:** Generated content may perpetuate harmful biases
- **Misuse Potential:** Techniques can be used for malicious purposes
- **Privacy Violations:** Steganographic systems may compromise user privacy
- **Regulatory Compliance:** Lack of frameworks for responsible use

TrojanStego exemplifies these concerns, as it can embed secrets directly into LLM outputs, potentially enabling data exfiltration and other malicious activities.

6.5.7 White-box vs. Black-box Trade-offs. The choice between white-box and black-box approaches involves fundamental trade-offs:

Aspect	**White-box**	**Black-box**	**Hybrid**
Security	High (95-99%)	Medium (79-91%)	Medium-High (90-95%)
Accessibility	Low	High	Medium
Capacity	High (1.1-5.98 bpt)	Medium (5.37 bpw)	Medium (2.0-4.0 bpt)
Imperceptibility	High (PPL: 3-8)	Low (PPL: 168-363)	Medium (PPL: 50-150)
Deployment	Difficult	Easy	Moderate

Table 9. Trade-offs between white-box, black-box, and hybrid approaches

6.5.8 Computational and Resource Constraints. Performance optimization often conflicts with computational efficiency:

- **Computational Overhead:** Better results typically require more computational resources
- **Memory Requirements:** Large models and external knowledge increase memory needs

[Placeholder footnote]

- **Real-time Constraints:** Latency requirements may limit optimization options
- **Scalability Issues:** Performance may degrade with increased scale

UTF demonstrates this trade-off, showing a 5% drop in HellaSwag performance, while **FreStega** requires corpus access (100 samples) for optimal performance.

6.5.9 Unresolved Challenges and Future Needs. Several critical challenges remain inadequately addressed:

- **Provable Security:** Lack of theoretical foundations for security guarantees
- **Robustness:** Limited resilience to advanced attack methods
- **Standardization:** Absence of common evaluation frameworks
- **Ethical Frameworks:** Missing guidelines for responsible development and use
- **Cross-lingual Support:** Poor performance in non-English languages
- **Real-world Deployment:** Limited testing in actual deployment scenarios

6.5.10 Quantitative Impact Analysis. The following table provides a quantitative overview of the most significant 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 10. Quantified impact of key limitations and trade-offs

Understanding these limitations and trade-offs is essential for advancing the field and developing more robust, secure, and practical steganographic systems. Future research must address these challenges to enable widespread adoption and responsible use of LLM-based steganography.

Table 11. Summary of Results from Reviewed Papers

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
VAE-Stega: linguistic steganography based on va... [39]	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

Continued on next page

Table 11 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
General framework for reversible data hiding in... [44]	BERTBase	BookCorpus	BPW=0.5335 F1=0.9402 PPL=134.2199	non-explicit	pre-text	text
Co-stega: Collaborative linguistic stegano-graph... [18]	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... [7]	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

Continued on next page

[Placeholder footnote]

Table 11 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
Generative text steganography with large langua... [36]	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]
Meteor: Cryptographically secure steganography ... [14]	GPT-2	Hutter Prize, HTTP GET requests	GPT-2: 3.09 bits/token	non-explicit	tuning + pretext	text
Zero-shot generative linguistic steganography [19]	LLaMA2-Chat-7B (as the stegotext generator / QA model). GPT-2 (for NLS baseline and JSD evaluation)	IMDB, Twitter	PPL: 8.81. JSDfull: 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... [26]	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

Continued on next page

Table 11 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
A principled approach to natural language water... [13]	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
Context-aware linguistic steganography model ba... [6]	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... [22]	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... [35]	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

Continued on next page

[Placeholder footnote]

Table 11 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
Linguistic steganography: From symbolic space t... [41]	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
Natural language steganography by chatgpt [33]	[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... [27]	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... [16]	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]

Continued on next page

[Placeholder footnote]

Table 11 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
ALiSa: Acrostic linguistic steganography based ... [40]	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 DISCUSSION

This section provides a comprehensive discussion of the findings presented in the results section, synthesizing insights across all research questions and identifying implications for future research and practice.

7.1 Synthesis of Key Findings

The systematic review reveals a rapidly evolving field that has undergone significant transformation since 2023. The shift from white-box to black-box approaches represents a paradigm change toward more practical, real-world deployable steganographic systems. This evolution is driven by the increasing accessibility of large language models through APIs and the need for covert communication in censored environments.

7.2 Implications for Research and Practice

7.2.1 Methodological Implications. The findings suggest several important methodological considerations:

- **Standardization Need:** The lack of standardized evaluation metrics and benchmarks represents a critical barrier to progress. Future research should prioritize the development of common evaluation frameworks.
- **Evaluation Completeness:** The limited use of human evaluation (only 25% of studies) and robustness testing (40% missing) indicates a need for more comprehensive evaluation practices.
- **Reproducibility:** The variation in reporting standards and missing implementation details in many studies hampers reproducibility and comparison.

7.2.2 Practical Implications. For practitioners and developers:

- **Method Selection:** The choice between white-box and black-box methods should be based on security requirements vs. deployment constraints.
- **Capacity Planning:** The Psic Effect and capacity limitations in short texts should be carefully considered in system design.
- **Security Considerations:** The vulnerability to attacks (5-50% detection rate drops) requires robust defense mechanisms.

[Placeholder footnote]

7.3 Addressing the Psic Effect

The Perceptual-Statistical Imperceptibility Conflict emerges as the most significant challenge in the field. This fundamental trade-off between perceptual quality and statistical security affects 80% of studies and results in an average capacity loss of 1-2 bits per word. Future research should focus on:

- Developing techniques that minimize this trade-off
- Creating adaptive systems that balance both aspects dynamically
- Exploring novel approaches that decouple perceptual and statistical imperceptibility

7.4 The Role of Context and External Knowledge

The integration of external knowledge sources has proven crucial for enhancing both capacity and contextual relevance. However, this integration introduces new challenges:

- **Privacy Concerns:** External knowledge integration may compromise the privacy of the steganographic system
- **Computational Overhead:** The 5-15% increase in computational cost may limit real-time applications
- **Generalizability:** Domain-specific knowledge may not transfer well across different contexts

7.5 Ethical Considerations and Responsible Development

The review reveals a concerning gap in ethical considerations, with only 10% of studies addressing ethical implications. This represents a significant oversight given the potential for misuse in:

- Censorship evasion in authoritarian regimes
- Covert communication for malicious purposes
- Data exfiltration and information leakage
- Bias propagation through generated content

Future research must prioritize the development of ethical frameworks and responsible use guidelines.

7.6 Limitations of the Review

Several limitations of this systematic review should be acknowledged:

- **Incomplete Coverage:** 14 papers remained pending PDF acquisition, potentially missing important insights
- **Language Bias:** The focus on English-language publications may have excluded relevant non-English research
- **Recency Bias:** The rapid evolution of the field means some recent developments may not be fully captured
- **Quality Assessment:** The lack of formal quality assessment tools may have influenced the synthesis

7.7 Future Research Directions

Based on the synthesis of findings, several promising research directions emerge:

7.7.1 Technical Advancements.

- **Multimodal Steganography:** Integration with vision-language models for text-image combinations
- **Robust Defense Mechanisms:** Development of attack-resistant techniques
- **Provable Security:** Theoretical foundations for stronger security guarantees
- **Efficient Computation:** Reducing computational overhead for real-time applications

[Placeholder footnote]

7.7.2 *Methodological Improvements.*

- **Standardized Evaluation:** Development of common benchmarks and evaluation protocols
- **Human-Centered Design:** Greater emphasis on human evaluation and usability
- **Cross-Language Support:** Extension to non-English languages and cultural contexts
- **Real-World Testing:** Evaluation in actual deployment scenarios

7.7.3 *Ethical and Social Considerations.*

- **Ethical Frameworks:** Development of guidelines for responsible use
- **Bias Mitigation:** Techniques to prevent discrimination and bias propagation
- **Transparency:** Methods for detecting and auditing steganographic content
- **Regulatory Compliance:** Alignment with emerging AI regulations and standards

7.8 Conclusion

This systematic review has provided a comprehensive analysis of the current state of LLM-based steganography, revealing both significant progress and critical challenges. The field has evolved rapidly, with clear trends toward more practical and context-aware systems. However, fundamental limitations such as the Psic Effect, attack vulnerability, and ethical concerns remain inadequately addressed.

The findings suggest that future research should prioritize the development of standardized evaluation frameworks, robust defense mechanisms, and ethical guidelines. The integration of external knowledge sources shows promise but requires careful consideration of privacy and computational constraints. Most importantly, the field must address the ethical implications of these technologies to ensure their responsible development and deployment.

As LLMs continue to evolve and become more accessible, the field of linguistic steganography will likely see continued growth and innovation. The challenges identified in this review provide a roadmap for future research directions, while the opportunities suggest exciting possibilities for advancing both the technical capabilities and practical applications of these systems.

8 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

[Placeholder footnote]

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.

Table 12. Summary of Results from Reviewed Papers

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
VAE-Stega: linguistic steganography based on va... [39]	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... [44]	BERTBase	BookCorpus	BPW=0.5335 F1=0.9402 PPL=134.2199	non-explicit	pre-text	text
Co-stega: Collaborative linguistic steganograph... [18]	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

Continued on next page

[Placeholder footnote]

Table 12 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
Joint linguistic steganography with BERT masked... [7]	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... [36]	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]
Meteor: Cryptographically secure steganography ... [14]	GPT-2	Hutter Prize, HTTP GET requests	GPT-2: 3.09 bits/token	non-explicit	tuning + pre-text	text

Continued on next page

[Placeholder footnote]

Table 12 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
Zero-shot generative linguistic steganography [19]	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... [26]	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... [13]	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
Context-aware linguistic steganography model ba... [6]	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

Continued on next page

[Placeholder footnote]

Table 12 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
DeepTextMark: a deep learning-driven text water... [22]	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... [35]	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... [41]	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
Natural language steganography by chatgpt [33]	[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

Continued on next page

[Placeholder footnote]

Table 12 – continued from previous page

Paper	Llm	Dataset	Result	Context Aware	Categ Context	Representation Context
Natural language watermarking via paraphraser-b... [27]	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... [16]	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 ... [40]	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]

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] 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.

[4] 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).

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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).

[11] 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

[Placeholder footnote]

[12] 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

[13] 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.

[14] 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.

[15] 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

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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

[22] 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.

[23] 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.

[24] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. 2015. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology* 64 (08 2015). doi:10.1016/j.infsof.2015.03.007

[25] 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.

[26] 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.

[27] 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.

[28] 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

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

[30] 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

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

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

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

[34] 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).

[35] 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.

[36] 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.

[37] 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

[Placeholder footnote]

[38] 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

[39] 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.

[40] 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.

[41] 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.

[42] 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

[43] 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).

[44] 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.

[45] 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)*.

[Placeholder footnote]