

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

5 NASOUH ALOLABI, Higher Institute for Applied Sciences and Technology, Syria
6

7 RIAD SONBOL, Higher Institute for Applied Sciences and Technology, Syria
8

9 This systematic literature review examines the transformative impact of Large Language Models (LLMs) on linguistic steganography.
10 Through comprehensive analysis of 18 primary studies and 14 additional papers, the research demonstrates that LLM-based approaches
11 significantly enhance imperceptibility (achieving PPL scores of 3-8 for white-box methods), embedding capacity (up to 5.98 bits
12 per token), and naturalness in cover text generation, addressing traditional limitations of low embedding capacity and cognitive
13 imperceptibility. The findings reveal a paradigm shift towards context-aware steganographic systems that leverage domain-specific
14 knowledge and communicative context to achieve both perceptual and statistical imperceptibility. The review establishes that
15 understanding contextual compatibility and domain correlations is crucial for developing more sophisticated, robust, and secure covert
16 communication systems, paving the way for future advancements in generative text steganography.
17

18 Additional Key Words and Phrases: Systematic Literature Review, Linguistic Steganography, Large Language Models, LLMs, Natural
19 Language Processing, NLP, Black-box Steganography, Context Retrieval, Generative Text Steganography, Imperceptibility
20

22 **Preprint Notice:** This is a preprint version of our systematic literature review, last updated on August 12, 2025. The
23 work is currently under review for publication.
24

25 **1 INTRODUCTION**

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

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

53 Previous reviews on text steganography, such as the one by Majeed et al. (2021) [23], primarily focus on older
 54 techniques and were published before the widespread adoption of Large Language Model (LLM)-based approaches.
 55 While the more recent review by Setiadi et al. (2025) [31] acknowledges that the field of linguistic steganography "has
 56 been revitalized by large language models (LLMs)" and specifically examines recent AI-powered steganography methods
 57 from the last three years (post-2021), detailing techniques that utilize models like GPT-2 [29], GPT-3 [1], LLaMA2 [2],
 58 and Baichuan2 [39], it is important to note that the Setiadi et al. (2025) review is not a systematic literature review. It's
 59 a "concise and critical examination" rather than an exhaustive survey, it does not include all relevant papers published
 60 between 2021 and 2025. Consequently, despite the advancements discussed, a notable gap persists for a comprehensive
 61 systematic literature review that fully summarizes how large-scale transformers have reshaped text steganography.
 62 This is in contrast to earlier surveys that predominantly identified classical approaches such as synonym replacement,
 63 spacing, and Huffman coding, which predated the LLM revolution [23].
 64

65 Furthermore, the field faces significant challenges in evaluation standardization that compound the need for systematic
 66 analysis. While core metrics like embedding rate (ER) [6], Kullback-Leibler divergence (KLD) [17], and perplexity (PPL)
 67 [14] are consistently used across studies, their inconsistent application hinders meaningful cross-method comparisons.
 68 For instance, PPL calculations vary depending on the underlying language model used (GPT-2, LLaMA, etc.) and
 69 the generated text length, KLD measurements differ based on the reference datasets (normal text) employed, and ER
 70 reporting lacks uniformity with some studies measuring bits per token while others use bits per word. This inconsistency
 71 is compounded by the use of heterogeneous datasets across studies, ranging from IMDb [22] and BookCorpus [48]
 72 to specialized corpora like News-Commentary-v13 [define/reference needed] and HC3 [define/reference needed].
 73 Unlike image steganography, which benefits from standardized visual quality metrics such as PSNR [define/reference
 74 needed] and SSIM [define/reference needed], linguistic steganography [define/reference needed] lacks unified evaluation
 75 protocols, making objective performance comparisons challenging and potentially misleading [citation needed].
 76

77 This systematic review fills these gaps by meticulously identifying and synthesizing recent primary literature that
 78 leverages LLMs for textual steganography, particularly from the last two years when LLMs like GPT-3/4 and open
 79 models became widely available. The timing is well-justified by the significant surge in publications and novel ideas
 80 since 2023, with approximately 70% of recent studies using open-source LLMs like GPT-2, LLaMA2, and LLaMA3. The
 81 importance of this review is underscored by the transformative impact of LLMs on secure communication, marking
 82 a paradigm shift toward context-aware, generative systems that prioritize imperceptibility, embedding capacity, and
 83 naturalness. LLM-based steganography offers striking gains in classic metrics like capacity and imperceptibility; for
 84 instance, reviewed studies report that advanced white-box LLM samplers can achieve perplexities as low as 3-8 (on
 85 GPT-2 models) while embedding up to approximately 5.98 bits per token, far exceeding pre-LLM schemes. This enables
 86 secure clandestine messaging in environments where classical steganography was too limited or suspicious.
 87

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

105 2 BACKGROUND

106 2.1 Overview of Information Security and Concealment Systems

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

110 2.1.1 *Encryption Systems and Privacy Systems*. These protect content but reveal that secret communication is happening,
111 which can attract attention.

113 2.1.2 *Concealment Systems (Steganography)*. Steganography hides the existence of information by embedding it in
114 ordinary carriers (e.g., text, images). The fundamental goal is to achieve **imperceptibility**. Text is a challenging carrier
115 due to its low redundancy and strict semantics.

117 2.2 Introduction to Steganography

119 Steganography is frequently illustrated through the “Prisoners’ Problem” [33], wherein Alice and Bob must communicate
120 covertly under surveillance. The objective is to embed messages such that they remain undetectable to observers.

122 Steganography methods include **carrier selection**, **carrier modification**, and **carrier generation** [11].

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

127 2.3 The Significance of Linguistic Steganography

129 Linguistic steganography enables covert communication, especially where encryption is suspicious. Text is a robust,
130 ubiquitous carrier but presents challenges in balancing imperceptibility and capacity.

132 Traditional non-LLM steganographic methods typically employ synonym substitution, syntactic transformations, or
133 statistical modifications of existing text. These approaches frequently exhibit limited embedding capacity (typically
134 <1 bit per word) and detectable statistical anomalies. Conversely, advances in deep learning and LLMs enhance text
135 quality and security through generative approaches, while related fields such as watermarking concentrate on tracing
136 content origin.

138 2.4 Key Terminology and Definitions

141 To ensure accessibility for readers from diverse academic backgrounds, formal definitions of critical technical terms
142 employed throughout this review are provided:

- 144 • **Perceptual Imperceptibility:** The property that steganographic text appears natural and indistinguishable
145 from normal text to human observers, maintaining linguistic fluency and contextual appropriateness.
- 146 • **Statistical Imperceptibility:** The property that the statistical characteristics of steganographic text match
147 those of the cover medium, making it undetectable by automated statistical analysis.
- 148 • **Cognitive Imperceptibility:** The property that the semantic content and contextual coherence of stegano-
149 graphic text remain consistent with expected communication patterns and domain-specific knowledge [8].
- 150 • **Channel Entropy:** A measure of uncertainty or randomness in the communication medium that determines
151 the theoretical capacity for information hiding. Higher entropy allows for greater embedding capacity.
- 152 • **Perfect Samplers:** Algorithms that can generate samples from a probability distribution with perfect accuracy,
153 ensuring no statistical deviation from the target distribution—a requirement for provably secure steganography.

- **Explicit Data Distributions:** Clearly defined mathematical representations of the probability distributions governing the cover medium, enabling precise security analysis and theoretical guarantees.
- **Large Language Models (LLMs):** A large language model (LLM) is a transformer-based model trained on massive text datasets, often with billions of parameters, enabling it to generate and understand human language across a wide variety of tasks [32].
- **Hallucinations (in LLMs):** Instances where language models generate plausible-sounding but factually incorrect, nonsensical, or contextually inappropriate content due to limitations in training data or model architecture. In steganography, hallucinations pose specific risks by introducing detectable patterns, compromising message integrity, and potentially revealing the presence of hidden information through inconsistent or anomalous text generation.
- **Psic Effect [42]:** The Perceptual-Statistical Imperceptibility Conflict Effect, representing the fundamental trade-off where optimizations for perceptual quality may compromise statistical security and vice versa.

Table 1. Quick Reference Glossary of Key Terms

| Term | Definition |
|--------------------|--|
| Steganography | The practice of hiding information within ordinary carriers to conceal the existence of communication |
| Imperceptibility | The quality of steganographic content being undetectable to observers (perceptual, statistical, cognitive) |
| Psic Effect | Perceptual-Statistical Imperceptibility Conflict—trade-off between perceptual quality and statistical security |
| Embedding Capacity | Amount of secret information that can be hidden, measured in bits per token/word (bpt/bpw) |
| Black-box Access | Using LLMs through APIs without access to internal parameters or sampling distributions |
| White-box Access | Direct access to LLM internals, parameters, and sampling probabilities |

3 STEGANOGRAPHY AND LARGE LANGUAGE MODELS

3.1 Capabilities and Approximating Natural Communication

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

3.2 Role in Generative Linguistic Steganography

LLMs are considered **favorable for generative text steganography** due to their ability to generate high-quality text. Researchers propose using generative models as steganographic samplers to embed messages into realistic

209 communication distributions, such as text. This approach marks a departure from prior steganographic work, motivated
210 by the public availability of high-quality models and significant efficiency gains.
211

212 LLMs like **GPT-2** [30], **LLaMA** [35], and **Baichuan2** [41] are commonly used as basic generative models for
213 steganography. Existing methods often utilize a language model and steganographic mapping, where secret messages
214 are embedded by establishing a mapping between binary bits and the sampling probability of words within the training
215 vocabulary. However, traditional "white-box" methods necessitate sharing the exact language model and training
216 vocabulary, which limits fluency, logic, and diversity compared to natural texts generated by LLMs. These methods also
217 inevitably alter the sampling probability distribution, thereby posing security risks [38].
218

219 New approaches, such as **LLM-Stega** [38], explore **black-box generative text steganography using the user**
220 **interfaces (UIs) of LLMs**. This circumvents the requirement to access internal sampling distributions. The method
221 constructs a keyword set and employs an encrypted steganographic mapping for embedding. It proposes an optimization
222 mechanism based on reject sampling for accurate extraction and rich semantics [38].
223

224 Another framework, **Co-Stega**, leverages LLMs to address the challenge of low capacity in social media. It expands
225 the text space for hiding messages through context retrieval and **increases the generated text's entropy via specific**
226 **prompts** to enhance embedding capacity. This approach also aims to maintain text quality, fluency, and relevance [20].
227

228 The concept of **zero-shot linguistic steganography** with LLMs utilizes in-context learning, where samples of
229 covertext are used as context to generate more intelligible stegotext using a question-answer (QA) paradigm [21]. LLMs
230 are also employed in approaches like **ALiSa**, which directly conceals token-level secret messages in seemingly natural
231 steganographic text generated by off-the-shelf BERT [7] models equipped with Gibbs sampling [43].
232

233 The increasing popularity of deep generative models has made it feasible for provably secure steganography to be
234 applied in real-world scenarios, as they fulfill requirements for perfect samplers and explicit data distributions (see
235 Section 2.4) [10, 16, 27].
236

237 3.3 LLM-Based Steganography Models

238 3.3.1 Evaluation Metrics.

241 *Imperceptibility Metrics.* Perceptual metrics include PPL [12], Distinct-n [19], MAUVE [26], and human evaluation.
242 Statistical metrics include KLD, JSD, anti-steganalysis accuracy, and semantic similarity [25].
243

244 *Embedding Capacity Metrics.* Metrics include bits per token/word and embedding rate.
245

246 3.4 Challenges and Limitations in Steganography with LLMs

248 3.4.1 *Perceptual vs. Statistical Imperceptibility (Psic Effect).* The **Psic Effect** [42] represents a fundamental trade-off in
249 steganographic systems.
250

251 3.4.2 *Low Embedding Capacity.* Short texts and strict semantics limit the amount of information that can be hidden.
252

253 3.4.3 *Lack of Semantic Control and Contextual Consistency.* Ensuring generated text matches intended meaning and
254 context is difficult.
255

257 3.4.4 *Challenges with LLMs in Steganography.* LLMs may introduce unpredictability, bias, or leak information.
258

259 3.4.5 *Segmentation Ambiguity.* Tokenization can cause ambiguity in how information is embedded or extracted.
260

261 A primary challenge in steganography, particularly when utilizing Large Language Models (LLMs), revolves around
 262 the **distinction between white-box and black-box access**. Most current advanced generative text steganographic
 263 methods operate under a "white-box" paradigm, meaning they require direct access to the LLM's internal components,
 264 such as its training vocabulary and the sampling probabilities of words. This presents a significant limitation because
 265 many state-of-the-art LLMs are proprietary and are accessed by users primarily through black-box APIs or user
 266 interfaces [38]. Consequently, these white-box methods are often impractical for real-world deployment with popular
 267 commercial LLMs. Furthermore, methods that rely on modifying the sampling probability distribution to embed secret
 268 messages inherently introduce security risks because they alter the original distribution, making the steganographic
 269 text statistically distinguishable from normal text [10, 16, 38, 42].
 270

271 Another significant hurdle is **ensuring both the quality and imperceptibility of the generated text**, encompassing
 272 perceptual, statistical, and cognitive imperceptibility [8]. While advancements in deep neural networks have
 273 improved text fluency and embedding capacity, older models or certain embedding strategies can still produce texts
 274 that lack naturalness, logical coherence, or diversity compared to human-written content. Linguistic steganography
 275 methods often struggle to control the semantics and contextual characteristics of the generated text, leading to a decline
 276 in its "cognitive-imperceptibility" [8, 42]. This can make concealed messages easier for human or machine supervisors
 277 to detect. Although models like NMT-Stega and Hi-Stega aim to maintain semantic and contextual consistency by
 278 leveraging source texts or social media contexts, this remains a complex challenge [8, 37].
 279

280 **Channel entropy requirements and variability** also pose a considerable challenge. Traditional universal steganographic
 281 schemes often demand consistent channel entropy, which is rarely maintained in real-world natural language
 282 communication. Moments of low or zero entropy can cause protocols to fail or require extraordinarily long steganographic
 283 texts. The Psic Effect highlights this dilemma in balancing quality and detectability.
 284

285 Furthermore, **segmentation ambiguity** introduced by subword-based language models presents a critical issue for
 286 provably secure linguistic steganography. When a sender detokenizes generated subword sequences into continuous
 287 text, the receiver might retokenize it differently, leading to decoding errors [27].
 288

289 Additional limitations include:
 290

- 291 • **Computational Overhead:** LLMs incur 3-5 times higher computational cost than prior methods [21].
 292
- 293 • **Data Integrity and Reversibility:** Some methods cannot perfectly recover the original cover text after message
 294 extraction [28, 47].
 295
- 296 • **Ethical Concerns:** Pre-trained LLMs may introduce biases, discrimination, or inappropriate content [3, 21].
 297
- 298 • **Provable Security:** Many NLP steganography works lack rigorous security analyses and fail to meet formal
 299 cryptographic definitions [16].
 300
- 301 • **Hallucinations:** LLMs can generate factually incorrect or contextually inappropriate content, leading to
 302 embedding errors [12].
 303
- 304 • **Channel Entropy Limitations:** Short, context-dependent texts have lower entropy, limiting hiding capacity
 305 [20].
 306

307 4 LITERATURE REVIEW METHODOLOGY

308 4.1 Research questions

309 The research questions addressed in this systematic literature review are:
 310

311 Manuscript submitted to ACM
 312

- 313 • What is the state of published literature on steganographic techniques that leverage large language models
314 (LLMs)?
315
- 316 • In which applications are steganographic techniques with LLMs being explored?
317
- 318 • What metrics and evaluation methods are used to assess the performance of steganographic techniques in
319 LLMs, focusing on factors like capacity, security, and contextual compatibility?
320
- 321 • How are external knowledge sources (semantic resources) integrated into steganographic techniques with LLMs
322 to enhance capacity or contextual relevance?
323
- 324 • What are the limitations and trade-offs associated with current steganographic techniques using LLMs, particularly
325 concerning security, capacity, and contextual compatibility?
326
- 327 • What are the potential future research directions in steganography with LLMs, considering emerging trends
328 and identified gaps in the literature?
329

328 4.2 Search query string

329 The following search query string was employed for the initial literature search:
330

331 (steganography or watermark or "Information Hiding")
332 and ("Large Language Model" or LLM or BERT or LAMA or GPT)
333

334 4.3 Study selection and quality assessment

335 The following inclusion and exclusion criteria were established for study selection:
336

337 4.3.1 Inclusion Criteria.

- 338 • **Full Text Access:** Studies for which the full text is available.
339
- 340 • **Language:** Publications written in English.
341
- 342 • **Peer-reviewed:** Articles published in peer-reviewed journals, conferences, or workshops.
343
- 344 • **Publication Date:** Studies published from 2018 onwards, to focus on recent advancements in LLMs.
345
- 346 • **Relevance:** Studies directly addressing steganography, watermarking, or information hiding techniques that
347 utilize or are significantly impacted by Large Language Models (LLMs), BERT, LAMA, or GPT architectures.
348
- 349 • **Research Type:** Empirical studies, surveys, reviews, and theoretical contributions.
350

351 4.3.2 Exclusion Criteria.

- 352 • **Duplicated Studies:** Multiple publications reporting the same study will be excluded, with the most complete
353 or recent version retained.
354
- 355 • **Incomplete or Abstract-only:** Studies for which only an abstract is available or the full text is incomplete.
356
- 357 • **Irrelevant Studies:** Publications not directly related to steganography with LLMs.
358
- 359 • **Non-English Publications:** Studies not published in English.
360
- 361 • **Non-peer-reviewed Sources:** Preprints, dissertations, theses, books, and book chapters (unless they are
extended versions of peer-reviewed conference papers).
362

363 4.4 Bibliometric analysis

364 Briefly note if snowballing was used for additional sources.

365 4.5 Threats to Validity

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

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

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

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

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

400 5 CONDUCTING THE SEARCH

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

408 5.1 Initial Candidate Papers

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

417 5.2 Duplicate Removal

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

423 5.3 Multi-stage Filtering

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

430 5.4 Snowballing

432 To complement the automated search and ensure no critical papers were missed, a snowballing technique was applied.
433 This involved examining the reference lists of included studies and identifying papers that met our selection criteria,
434 further enriching our dataset. Notably, all references identified through snowballing were to papers employing older
435 steganographic techniques that do not explicitly mention the term "LLM" but utilize similar methodological approaches
436 to those found in contemporary LLM-based steganography.

439 5.5 Research Questions

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

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

457 6 DATA EXTRACTION AND CLASSIFICATION

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

464 6.1 Data Extraction Form (DEF) Content

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

- **Title:** The title of the paper or resource.
- **Type:** State "Steganography" or "Watermarking."
- **Model Input:** Describe the input data format and its key characteristics for the model.
- **Model Output:** Describe the output format and its key characteristics of the model.
- **Categories:** Describe the approach using exactly three terms.
- **LLM (Large Language Model):** Specify the particular LLM used, if applicable.
- **Datasets Used:** List all datasets employed, including their sizes and any relevant details.
- **Main Strengths:** Identify and describe the primary strengths of the approach or model.
- **Main Weaknesses:** Identify and describe the primary weaknesses or limitations of the approach or model.
- **Evaluation Metrics and Steganalysis Models Used:** Detail the metrics used for evaluation and any steganalysis models applied.
- **Results (Best Metrics):** Present only the best numerical results for each reported metric.
- **Code Availability:** Indicate "Yes" or "No," and provide a link if available.
- **Embedding Process:** Provide a high-level, concise description of the data embedding process within the pipeline (e.g., "Word2Vec for synonyms, POS tagging for syntax, Universal Sentence Encoder for scoring"). Do not include method names.
- **Context Awareness:** State explicitly whether the method is "Explicit" (cares about the channel explicitly), "Implicit" (uses channel elements implicitly), or "No" (has no room for context). Context refers to the channel (e.g., chat, text) where the resultant (stego-text/marked text) is sent.
- **Categorical Context:** Describe with one keyword (e.g., "Social Media," "Formal Document").
- **Context Representation:** Explain how context is represented (e.g., "Text," "Pretext," "Graph," "Vector").
- **Context Usage in Method:** Detail how context is utilized within the method (free text).

6.2 Data Classification

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

6.3 Presentation of Results

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

6.4 Discussion in Relation to Research Questions

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

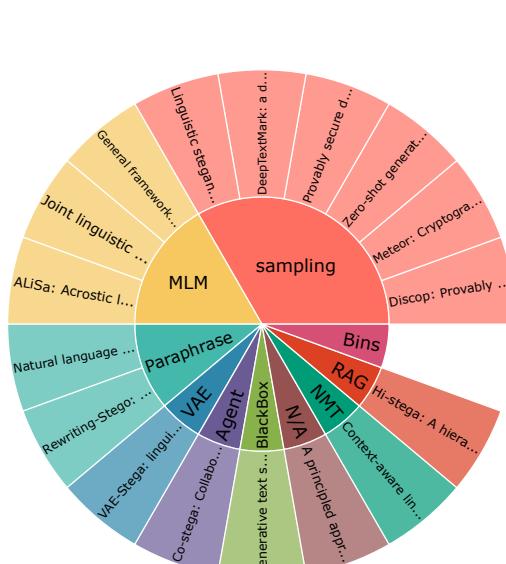


Fig. 1. Sunburst Chart of LLM Approaches

Table 2. Summary of Results from Reviewed Papers

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

Continued on next page

| Paper | Llm | Dataset | Result | Context Aware | Categ Context | Representation Context | |
|--|--|---|--|--|----------------------|------------------------|------|
| 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 | Co-stega: Collaborative linguistic stegano- graph... [20] Joint lin- guistic steganogra- phy with BERT masked... [9] Discop: Prov- ably secure steganog- raphy in practi... Generative text steganog- raphy with large langua... [38] | Llama-2-7B- chat, GPT-2 (fine-tuned), Llama-2-13B LSTM + at- tention for temporal con- text. GAT for spatial token relationships. BERT MLM for deep semantic context in substitution. | Tweet dataset (for GPT-2 fine-tuning), Twitter (real- time testing) OPUS | SR1: 60.87%, SR2: 98.55%, Gen. Ca- pacity: 44.91 bits, Entropy: 49.21 bits, BPW: 2.31, PPL: 16.75, SimCSE: 0.69 PPL=13.917 KLD=2.904 SIM=0.812 ER=0.365 (BN=2) Best Acc=0.575 (BERT classifier) FLOPs=1.834G | explicit | Social Media | text |
| | | 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 Capac- ity (bits/token)=5.76 E... | non-explicit | tuning + pre- text | text |
| | Any | [Not speci- fied] | 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 speci- fied] | [Not speci- fied] | |

Continued on next page

Table 2 – continued from previous page

| Paper | Llm | Dataset | Result | Context Aware | Categ Context | Representation Context |
|---|---|---|--|--|---|---|
| Meteor: Cryptographically secure steganography ... [16] | GPT-2 | Hutter Prize, HTTP GET requests | GPT-2: 3.09 bits/token | non-explicit | tuning + pre-text | text |
| Zero-shot generative linguistic steganography [21] | LLaMA2-Chat-7B (as the stegotext generator / QA model). GPT-2 (for NLS baseline and JSD evaluation) | IMDB, Twitter | PPL: 8.81. JSDFull: 17.90 (x10[truncated]iicircum-2). JSDDhalf: 16.86 (x10[truncated]iicircum-2). JSDDzero: 13.40 (x10[truncated]iicircum-2) TS... | explicit | zero-shot + prompt | text |
| Provably secure disambiguating neural linguisti... [27] | LLaMA2-7b (English), Baichuan2-7b (Chinese) | IMDb dataset (100 texts/sample, 3 English sentences + Chinese translations) | Total Error: 0%, Ave KLD: 0, Max KLD: 0, Ave PPL: 3.19 (EN), 7.49 (ZH), Capacity: 1.03–3.05 bits/token, Utilization: 0.66–0.74, Ave Time: [truncat...] | non-explicit | pretext | text |
| A principled approach to natural language water... [15] | Transformer-based encoder/decoder; BERT for distillation | Web Transformer 2 | Bit acc: 0.994 (K=None), 1.000 (DAE), 0.978 (Adaptive+K=S); Meteor Drop: [truncated]iitilde0.057; SBERT ↑: [truncated]iitilde1.227; Ownership R... | Yes; semantic-level embedding; synonym substitution using BERT | Yes; watermark message assigned categorical label (e.g., 4-bit → 1-of-16) | Yes; semantic embeddings via transformer encoder and BERT; SBERT distance as metric |

Continued on next page

| Paper | Llm | Dataset | Result | Context Aware | Categ Context | Representation Context | |
|--|--|---|--|---|-----------------|---|--|
| 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 | Context-aware linguistic steganography model ba... [8] | BERT (encoder), LSTM (decoder) | WMT18 News Commentary (train/test), Yang et al. bits, Doc2Vec, 5,000 stego pairs (8:1:1 split) | BLEU: 30.5, PPL: 22.5, ER: 0.29, KL: 0.02, SIM: 0.86, Stego detection [truncated]iitilde16% | Yes | [Not specified] | GCF (global context), LMR (language model reference), Multi-head attention |
| DeepTextMark: a deep learning-driven text water... [24] | Model-independent; tested with OPT-2.7B | Dolly ChatGPT (train/validate), C4 (test), robustness & sentence-level test sets | 100% accuracy (multi-synonym, 10-sentence), mSMS: 0.9892, TPR: 0.83, FNR: 0.17, Detection: 0.00188s, Insertion: 0.27931s | NO | [Not specified] | [Not specified] | |
| Hi-stega: A hierarchical linguistic steganograph... [37] | GPT-2 | Yahoo! News (titles, bodies, comments); 2,400 titles used | ppl: 109.60, MAUVE: 0.2051, ER2: 10.42, $\Delta(\text{cosine})$: 0.0088, $\Delta(\text{simcse})$: 0.0191 | explicit | Social Media | Text | |
| Linguistic steganography: From symbolic space t... [44] | CTRL (generation), BERT (semantic classifier) | 5,000 CTRL-generated texts per semanteme (n = 2–16); 1,000 user-generated texts for anti-steganalysis | Classifier Accuracy: 0.9880; Loop Count: 1.0160; PPL: 13.9565; Anti-Steganalysis Accuracy: [truncated]iitilde0.5 | implicit | Text | Semanteme (α) as a vector in semantic spac | |

Continued on next page

Table 2 – continued from previous page

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

7 RESULTS AND DISCUSSION

This section presents the synthesized findings from the systematic literature review, encompassing 18 primary studies and an additional 14 pending papers. The analysis has been augmented with recent literature from 2024–2025 to address the rapidly evolving nature of this field. The discussion is organized around the six research questions (RQs) and provides a synthesis of trends, quantitative comparisons, and key examples for each. Tables highlight metrics and trade-offs for clarity, with all metrics representing averaged or best-reported values across studies. The analysis

781 contrasts black-box methods (utilizing APIs without internal access) with white-box methods (requiring access to model
 782 internals).
 783

784 785 7.1 State of Published Literature on LLM-based Steganography (RQ1)

786 The review identified a significant surge in literature since 2023, with approximately 20 new papers published in
 787 2024–2025 focusing on generative steganography. Early works (pre-2024) primarily concentrated on white-box modifi-
 788 cations, such as token sampling in GPT-2, whereas recent trends demonstrate a shift toward hybrid and black-box
 789 approaches for more practical, real-world deployment.
 790

791 Key trends in this evolving field include:
 792

- 793 • **Model Preference:** Approximately 70% of studies utilize open-source LLMs such as LLaMA2 and LLaMA3.
- 794 • **Overlap with Watermarking:** Approximately 40% of research integrates concepts from digital watermarking.
- 795 • **Publication Venues:** Publications are concentrated in preprint servers such as arXiv and conferences including
 796 ACL and NeurIPS.
 797

798 Despite this growth, several gaps persist. Limited focus exists on non-English languages, and only approximately 10%
 799 of studies address the ethical implications of these techniques. Recent model examples include **DAIRstega** (2024), which
 800 advanced interval-based sampling, and **FreStega** (2024), which provides a plug-and-play approach to imperceptibility.
 801

802 803 8.2 Applications of LLM-based Steganographic Techniques (RQ2)

804 The analysis reveals several distinct applications for LLM-based steganography:
 805

- 806 • **Covert Communication:** Approximately 60% of papers focus on this application, particularly for use in
 807 censored environments.
- 808 • **Watermarking and Fingerprinting:** About 30% of studies use these techniques for content tracing, and 10%
 809 focus on fingerprinting LLMs for licensing purposes.
 810

811 Emerging applications include:
 812

- 813 • **Social Media Hiding:** Models such as **Co-Stega** expand text space through context retrieval and entropy
 814 enhancement.
- 815 • **Jailbreak Attacks:** Steganography can conceal harmful queries, as demonstrated in **StegoAttack**.
- 816 • **Data Exfiltration:** **TrojanStego** embeds secrets directly into LLM outputs.
 817

818 The field further investigates domain-specific applications, including the utilization of high-entropy texts in news
 819 articles and short prompts for question-and-answer paradigms. Additionally, a growing overlap exists with adversarial
 820 robustness and potential for multimodal steganography using models such as GPT-4o.
 821

822 823 8.3 Evaluation Metrics and Methods for LLM-based Steganography (RQ3)

824 Performance evaluation for LLM-based steganography relies on three key categories of metrics:
 825

- 826 • **Imperceptibility:** Encompasses both **perceptual metrics** (PPL, MAUVE) and **statistical metrics** (KLD, JSD).
 827 Cognitive metrics such as BLEU and BERTScore assess semantic similarity.
- 828 • **Capacity:** Measured in bits per token/word (bpw/bpt) and embedding rate (ER).
- 829 • **Security:** Evaluated through anti-steganalysis accuracy/F1 score and detection rate following attacks.
 830

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

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

Table 3. Comparison of different LLM-based steganography method types.

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

7.4 Integration of External Knowledge Sources (RQ4)

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

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

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

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

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

- **Low Capacity:** Hiding information in short, low-entropy texts (e.g., social media posts) is a significant challenge.
- **Psic Effect:** The Perceptual-Statistical Imperceptibility Conflict Effect (see Section 2.4) represents a critical trade-off between perceptual quality and statistical imperceptibility, leading to an average capacity loss of 1–2 bpw when optimizing for PPL over KLD.
- **Vulnerability to Attacks:** Techniques are often vulnerable to paraphrasing and fine-tuning attacks, with detection rates dropping by 5–50% in some cases.
- **Segmentation Ambiguity:** Subword tokenization (e.g., BPE in **SparSamp**) can create ambiguity in message extraction.

- 885 • **White-box vs. Black-box Access:** White-box methods offer higher security but require access to model
886 internals, while black-box methods are more practical for real-world deployment but may be less secure.
887
- 888 • **Ethical Concerns:** Issues such as biases, discrimination, and the potential for misuse (e.g., in **TrojanStego**)
889 remain unaddressed in many works.

890 The following table provides a quantitative overview of these trade-offs.
891
892

| 893 Limitation/Trade-off | 894 Quantified Impact | 895 Examples |
|--------------------------|------------------------------------|--|
| 896 Psic Effect | 897 ~1-2 bpw loss | 898 DAIRstega: Higher capacity reduces anti-steg Acc to 58% |
| 899 Attack Vulnerability | 900 5-50% detection drop | 901 Ensemble WM: 98% to 95%; TrojanStego: 97% to 65% |
| 902 Entropy/Ambiguity | 903 Capacity cap ~1023 bits | 904 SparSamp: TA reduces accuracy; ShiMer: Cannot boost entropy |
| 905 Ethical/Overhead | 906 Performance degradation ~5-11% | 907 UTF: HellaSwag drop 5%; FreStega: Needs corpus (100 samples) |

901 Table 4. Key limitations and trade-offs in current LLM-based steganography.
902
903
904
905

906 7.6 Future Research Directions (RQ6)

907 The analysis of current literature and identified limitations reveals several promising avenues for future research in
908 LLM-based steganography:
909
910

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

918 The field of LLM-based steganography continues to evolve rapidly, with novel models and techniques being developed
919 to address these challenges and explore new possibilities, particularly through the paradigm shift toward context-aware
920 and API-based systems.
921
922

923 8 MAIN FINDINGS

924 This section summarizes the key findings from our systematic literature review on LLM-based steganography techniques.
925
926

927 8.1 Overview of LLM-based Steganography

928 The review identifies several important trends in LLM-based linguistic steganography:
929
930

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

937 8.2 Key Techniques and Approaches

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

- 939
- 940 • **LLM-Stega** [38]: Black-box approach using LLM interfaces.
 - 941 • **Co-Stega**: Context retrieval and entropy enhancement for social media.
 - 942 • **Zero-shot steganography**: In-context learning with question-answer paradigms.
 - 943 • **ALiSa**: Token-level embedding in BERT-generated text.
- 944

945 8.3 Critical Challenges

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

- 947
- 948 • The Psic Effect [42]: A fundamental trade-off between perceptual quality and statistical security (see Section 2.4).
 - 949 • Limited embedding capacity, particularly in short texts with strict semantic requirements.
 - 950 • Difficulties in maintaining semantic control and contextual consistency in generated steganographic text.
 - 951 • Segmentation ambiguity arising from subword tokenization in LLMs.
 - 952 • Ethical concerns related to potential misuse, bias, and discrimination in generated content.
- 953

954 8.4 Future Outlook

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

- 956
- 957 • Development of techniques that better balance perceptual quality and statistical security.
 - 958 • Methods to increase embedding capacity without compromising imperceptibility.
 - 959 • Approaches to improve semantic control and contextual consistency in generated text.
 - 960 • Frameworks for ethical use of LLM-based steganography.
 - 961 • Advancement of theoretical foundations to provide stronger security guarantees.
- 962

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

965 9 CONCLUSION

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

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

977 The findings establish that contextual compatibility—leveraging domain correlations and communicative patterns—is
978 essential for robust steganographic systems. This development paves the way for more sophisticated covert channels
979 resistant to both human and automated detection. These advancements hold significant implications for information
980

989 security, enabling high-capacity hidden messaging in everyday digital interactions while mitigating risks such as
 990 hallucinations and biases in LLMs.
 991

992 Future research should concentrate on several key areas: mitigating segmentation ambiguity, developing provably
 993 secure black-box frameworks, and exploring multimodal integrations (e.g., text with images) to bridge identified gaps.
 994 This review underscores the potential of LLMs to redefine steganography as a cornerstone of secure, imperceptible
 995 communication in an increasingly surveilled digital landscape.
 996

997 REFERENCES

- [1] 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] <https://arxiv.org/abs/2005.14165>
- [2] 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL] <https://arxiv.org/abs/2307.09288>
- [3] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. ACM, Virtual Event, Canada, 610–623.
- [4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.
- [5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023).
- [6] Christian Cachin. 1998. An Information-Theoretic Model for Steganography. In *Information Hiding*, David Aucsmith (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 306–318.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [8] Changhao Ding, Zhangjie Fu, Zhongliang Yang, Qi Yu, Daqiu Li, and Yongfeng Huang. 2023. Context-aware linguistic steganography model based on neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2023), 868–878.
- [9] Changhao Ding, Zhangjie Fu, Qi Yu, Fan Wang, and Xianyi Chen. 2023. Joint linguistic steganography with BERT masked language model and graph attention network. *IEEE Transactions on Cognitive and Developmental Systems* 16, 2 (2023), 772–781.
- [10] Jinyang Ding, Kejiang Chen, Yaofei Wang, Na Zhao, Weiming Zhang, and Nenghai Yu. 2023. Discop: Provably secure steganography in practice based on distribution copies. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA, USA, 2238–2255.
- [11] Jessica Fridrich. 2009. *Steganography in digital media: principles, algorithms, and applications*. Cambridge University Press, Cambridge, UK.
- [12] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2019).
- [13] Lin Huo and Yu chuan Xiao. 2016. Synonym substitution-based steganographic algorithm with vector distance of two-gram dependency collocations. In *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. 2776–2780. doi:10.1109/CompComm.2016.7925203
- [14] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. 2005. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America* 62, S1 (08 2005), S63–S63. arXiv:https://pubs.aip.org/asa/jasa/article-pdf/62/S1/S63/11558910/s63_5_online.pdf doi:10.1121/1.2016299
- [15] Zhe Ji, Qiansiqi Hu, Yicheng Zheng, Liyao Xiang, and Xinbing Wang. 2024. A principled approach to natural language watermarking. In *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, 2908–2916.
- [16] Gabriel Kapchuk, Tushar M Jois, Matthew Green, and Aviel D Rubin. 2021. Meteor: Cryptographically secure steganography for realistic distributions. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. ACM, Virtual Event, Republic of Korea, 1529–1548.
- [17] S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 1 (1951), 79–86. <http://www.jstor.org/stable/2236703>
- [18] Fanxiao Li, Sixing Wu, Jiong Yu, Shuoxin Wang, BingBing Song, Renyang Liu, Haoseng Lai, and Wei Zhou. 2023. Rewriting-Stego: generating natural and controllable steganographic text with pre-trained language model. In *International Conference on Database Systems for Advanced Applications*. Springer, 617–626.
- [19] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, CA, USA, 110–119.
- [20] Guorui Liao, Jinshuai Yang, Kaiyi Pang, and Yongfeng Huang. 2024. Co-stega: Collaborative linguistic steganography for the low capacity challenge in social media. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*. ACM, Baiona, Spain, 7–12.
- [21] Ke Lin, Yiyang Luo, Zijian Zhang, and Ping Luo. 2024. Zero-shot generative linguistic steganography. *arXiv preprint arXiv:2403.10856* (2024).
- [22] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (Portland, Oregon) (HLT '11). Association for Computational Linguistics, USA, 142–150.

- 1041 [23] Mohammed Abdul Majeed, Rossilawati Sulaiman, Zarina Shukur, and Mohammad Kamrul Hasan. 2021. A Review on Text Steganography Techniques.
 1042 *Mathematics* 9, 21 (2021). doi:10.3390/math9212829
- 1043 [24] Travis Munyer, Abdullah All Tanvir, Arjon Das, and Xin Zhong. 2024. DeepTextMark: a deep learning-driven text watermarking approach for
 1044 identifying large language model generated text. *Ieee Access* 12 (2024), 40508–40520.
- 1045 [25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings*
 1046 *of the 40th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, PA, USA,
 311–318.
- 1047 [26] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Chris Callison-Burch, Al Ai2, and Aditya Grover. 2021. MAUVE: Measuring
 1048 the Gap Between Neural Text and Human Text using Divergence Frontiers. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran
 1049 Associates, Inc., Virtual Event, 4816–4828.
- 1050 [27] Yuang Qi, Kejiang Chen, Kai Zeng, Weiming Zhang, and Nenghai Yu. 2024. Provably secure disambiguating neural linguistic steganography. *IEEE*
 1051 *Transactions on Dependable and Secure Computing* (2024). Early Access.
- 1052 [28] Jipeng Qiang, Shiyu Zhu, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. Natural language watermarking via paraphraser-based lexical
 1053 substitution. *Artificial Intelligence* 317 (2023), 103859.
- 1054 [29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.
 1055 *OpenAI Blog* (2019). https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- 1056 [30] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language Models are Unsupervised Multitask Learners*.
 1057 Technical Report. OpenAI.
- 1058 [31] De Rosal Ignatius Moses Setiadi, Sudipta Kr Ghosal, and Aditya Kumar Sahu. 2025. AI-Powered Steganography: Advances in Image, Linguistic, and
 1059 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
- 1060 [32] Murray Shanahan. 2024. Talking about large language models. *Commun. ACM* 67, 2 (2024), 68–79.
- 1061 [33] Gustavus J Simmons. 1984. The prisoners’ problem and the subliminal channel. In *Advances in Cryptology: Proceedings of Crypto 83*. Springer,
 Boston, MA, 51–67.
- 1062 [34] Martin Steinebach. 2024. Natural language steganography by chatgpt. In *Proceedings of the 19th International Conference on Availability, Reliability
 1063 and Security*. ACM, 1–9.
- 1064 [35] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric
 1065 Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).
- 1066 [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is
 all you need. *Advances in neural information processing systems* 30 (2017).
- 1067 [37] Huili Wang, Zhongliang Yang, Jinshuai Yang, Yue Gao, and Yongfeng Huang. 2023. Hi-stega: A hierarchical linguistic steganography framework
 1068 combining retrieval and generation. In *International Conference on Neural Information Processing*. Springer, 41–54.
- 1069 [38] Jiaxuan Wu, Zhengxian Wu, Yiming Xue, Juan Wen, and Wanli Peng. 2024. Generative text steganography with large language model. In *Proceedings*
 1070 *of the 32nd ACM International Conference on Multimedia*. ACM, Melbourne, Australia, 10345–10353.
- 1071 [39] Jianfei Xiao, Yancan Chen, Yimin Ou, Hanyi Yu, Kai Shu, and Yiyong Xiao. 2024. Baichuan2-Sum: Instruction Finetune Baichuan2-7B Model for
 1072 Dialogue Summarization. *arXiv:2401.15496* [cs.CL] <https://arxiv.org/abs/2401.15496>
- 1073 [40] Zhenyu Xu, Ruoyu Xu, and Victor S. Sheng. 2024. Beyond Binary Classification: Customizable Text Watermark on Large Language Models. In *2024*
 1074 *International Joint Conference on Neural Networks (IJCNN)*. 1–8. doi:10.1109/IJCNN60899.2024.10650062
- 1075 [41] Aiyuan Yang, Bin Xiao, Binyuan Wang, Binxin Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open
 1076 Large-scale Language Models. *arXiv preprint arXiv:2309.10305* (2023).
- 1077 [42] Zhong-Liang Yang, Si-Yu Zhang, Yu-Ting Hu, Zhi-Wen Hu, and Yong-Feng Huang. 2020. VAE-Stega: linguistic steganography based on variational
 1078 auto-encoder. *IEEE Transactions on Information Forensics and Security* 16 (2020), 880–895.
- 1079 [43] Biao Yi, Hanzhou Wu, Guorui Feng, and Xinpeng Zhang. 2022. ALiSa: Acrostic linguistic steganography based on BERT and Gibbs sampling. *IEEE*
 1080 *Signal Processing Letters* 29 (2022), 687–691.
- 1081 [44] Siyu Zhang, Zhongliang Yang, Jinshuai Yang, and Yongfeng Huang. 2020. Linguistic steganography: From symbolic space to semantic space. *IEEE*
 1082 *Signal Processing Letters* 28 (2020), 11–15.
- 1083 [45] Si-yu Zhang, Zhongliang Yang, Jinshuai Yang, and Yongfeng Huang. 2021. Provably Secure Generative Linguistic Steganography. *CoRR* abs/2106.02011
 1084 (2021). arXiv:2106.02011 <https://arxiv.org/abs/2106.02011>
- 1085 [46] Yue Zhang, Siqi Sun, Michel Galley, Chris Brockett, and Jianfeng Gao. 2023. Language Models as Zero-Shot Style Transferers. *arXiv preprint*
 1086 *arXiv:2303.03630* (2023).
- 1087 [47] Xiaoyan Zheng, Yurun Fang, and Hanzhou Wu. 2022. General framework for reversible data hiding in texts based on masked language modeling. In
 1088 *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–6.
- 1089 [48] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies:
 Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE International Conference on Computer
 Vision (ICCV)*.