

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 26 primary studies (6 pending acquisition), 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 hides secrets inside ordinary sentences—an exploit that looks trivial until one remembers how
28 little redundancy natural language actually contains [16, 39]. A single awkward synonym, a statistically rare clause,
29 or an out-of-place idiom is enough to alert an automated sentry. Classic tricks—swap a word here, bend the syntax
30 there—carry so few bits and leave such distinctive fingerprints that modern steganalysis routinely catches them [13].
31

32 Large language models change the game. Their uncanny fluency lets them spin entire documents that read like
33 human prose yet obey an adversarial agenda: every plausible continuation is also a potential codeword. The resulting
34 arms race has already produced generative schemes that write stego text from scratch [10, 36, 39, 42], rewriting engines
35 that paraphrase existing covers [18], black-box pipelines that treat the model as an opaque API [32, 36], zero-shot
36 protocols driven only by crafty prompting [20], collaborative frameworks that mine social context for extra entropy
37 [19, 35], and even constructions with provable indistinguishability guarantees [10, 16].
38

40 None of these victories is absolute. Push the embedding rate and the text begins to creak; optimize for statistical
41 stealth and the throughput collapses—the so-called “Psic effect” [39]. Segmentation ambiguities, computational overhead,
42 and the absence of shared benchmarks still slow progress. This survey dissects the advances, catalogs the open wounds,
43 and maps the territory that remains to be claimed.
44

45 Previous reviews on text steganography, such as the one by Majeed et al. (2021) [22], primarily focus on older
46 techniques and were published before the widespread adoption of Large Language Model (LLM)-based approaches.
47 While the more recent review by Setiadi et al. (2025) [29] acknowledges that the field of linguistic steganography “has
48 been revitalized by large language models (LLMs)” and specifically examines recent AI-powered steganography methods
49

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.

from the last three years (post-2021), detailing techniques that utilize models like GPT-2 [27], GPT-3 [1], LLaMA2 [2], and Baichuan2 [37], it is important to note that the Setiadi et al. (2025) review is not a systematic literature review. It's a "concise and critical examination" rather than an exhaustive survey, it does not include all relevant papers published between 2021 and 2025. Consequently, despite the advancements discussed, a notable gap persists for a comprehensive systematic literature review that fully summarizes how large-scale transformers have reshaped text steganography. This is in contrast to earlier surveys that predominantly identified classical approaches such as synonym replacement, spacing, and Huffman coding, which predated the LLM revolution [22].

Furthermore, the field faces significant challenges in evaluation standardization that compound the need for systematic analysis. While core metrics like embedding rate (ER) [6], Kullback-Leibler divergence (KLD) [17], and perplexity (PPL) [14] are consistently used across studies, their inconsistent application hinders meaningful cross-method comparisons. For instance, PPL calculations vary depending on the underlying language model used (GPT-2, LLaMA, etc.) and the generated text length, KLD measurements differ based on the reference datasets (normal text) employed, and ER reporting lacks uniformity with some studies measuring bits per token while others use bits per word. This inconsistency is compounded by the use of heterogeneous datasets across studies, ranging from IMDb [21] and BookCorpus [45] to specialized corpora like News-Commentary-v13 [define/reference needed] and HC3 [define/reference needed]. Unlike image steganography, which benefits from standardized visual quality metrics such as PSNR [define/reference needed] and SSIM [define/reference needed], linguistic steganography [define/reference needed] lacks unified evaluation protocols, making objective performance comparisons challenging and potentially misleading [citation needed].

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

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

[Placeholder footnote]

105 2 BACKGROUND

106 Information security systems broadly encompass **encryption**, **privacy**, and **concealment**, the last of which—known as
107 **steganography**—is the focus of this review. While encryption and privacy protect message content, they do not conceal
108 the existence of communication, which may itself arouse suspicion. Steganography instead prioritizes **imperceptibility**:
109 embedding information into ordinary carriers (e.g., images or text) so that hidden messages remain unnoticed.
110

111 Text is a particularly challenging carrier due to its low redundancy and strict semantic constraints. The classical
112 “Prisoners’ Problem” [31] illustrates the goal: two parties, Alice and Bob, must exchange hidden information without
113 alerting a watchful adversary.

114 Textual steganography methods are typically divided into **format-based** approaches, which exploit layout or
115 structural features, and **content-based** approaches, which modify linguistic form. Within the latter, early techniques
116 such as **synonym substitution** embed bits by altering lexical choices, but suffer from low capacity and high detectability.
117 More formally, **linguistic steganography** refers to concealing information in natural language by modifying or
118 generating text while preserving fluency and meaning [11].
119

120 Traditional linguistic approaches offer limited embedding capacity and often leave statistical artifacts. Advances in
121 deep learning and **Large Language Models (LLMs)** now enable generative methods that achieve higher text quality
122 and more secure embedding. Evaluating such systems requires several dimensions of imperceptibility: **perceptual**
123 (human naturalness), **statistical** (distributional similarity to natural text), and **cognitive** (semantic and contextual
124 fidelity) [8].
125

126 A deeper theoretical perspective introduces **channel entropy**, which quantifies the information-carrying capacity
127 of a given communication channel. Entropy sets the upper bound for embedding rates: higher entropy allows more
128 hidden information without detection, while lower entropy restricts capacity. Achieving this bound securely requires
129 **perfect samplers**, which can generate text indistinguishable from genuine distributional samples. These concepts
130 underpin the design of provably secure steganographic systems.
131

132 However, LLMs [30] introduce new challenges. Their tendency toward **hallucinations** can create detectable artifacts,
133 highlighting the **Psic Effect** (Perceptual-Statistical Imperceptibility Conflict) [39], where optimizing for perceptual
134 fluency may undermine statistical security. Model access further shapes practical steganography: with **black-box access**
135 (e.g., commercial APIs), developers gain scalability and ease of use but face limited control and reduced transparency. In
136 contrast, **white-box access** enables fine-grained control over parameters and sampling, supporting stronger security
137 guarantees, but requires costly resources and raises deployment barriers. This trade-off is central to evaluating the
138 robustness and applicability of modern linguistic steganography.
139

140 3 STEGANOGRAPHY AND LARGE LANGUAGE MODELS

141 3.1 Capabilities and Approximating Natural Communication

142 Large Language Models (LLMs) are autoregressive, generative systems based on the Transformer architecture [34] that
143 approximate high-dimensional distributions over natural-language sequences [16][28]. Given a prefix, an LLM emits a
144 probability vector over the vocabulary; the next token is sampled from this vector and appended to the prefix, and
145 the process repeats until a stopping criterion is met. During pre-training, billions of parameters are tuned on large
146 web corpora so that the model’s predictive distribution converges to the empirical distribution of the data [4]. As a
147 consequence, modern LLMs routinely produce text whose fluency, coherence and style are indistinguishable from
148

157 human writing [5]. The learned latent representations capture stylistic and semantic regularities that generalize across
 158 domains, enabling applications requiring nuanced linguistic mimicry [43].
 159

160 161 3.2 Role in Generative Linguistic Steganography

162 LLMs are considered **favorable for generative text steganography** due to their ability to generate high-quality
 163 text. Researchers propose using generative models as steganographic samplers to embed messages into realistic
 164 communication distributions, such as text. This approach marks a departure from prior steganographic work, motivated
 165 by the public availability of high-quality models and significant efficiency gains.
 166

167 LLMs like **GPT-2** [28], **LLaMA** [33], and **Baichuan2** [38] are commonly used as basic generative models for
 168 steganography. Existing methods often utilize a language model and steganographic mapping, where secret messages
 169 are embedded by establishing a mapping between binary bits and the sampling probability of words within the training
 170 vocabulary. However, traditional "white-box" methods necessitate sharing the exact language model and training
 171 vocabulary, which limits fluency, logic, and diversity compared to natural texts generated by LLMs. These methods also
 172 inevitably alter the sampling probability distribution, thereby posing security risks [36].
 173

174 New approaches, such as **LLM-Stega** [36], explore **black-box generative text steganography using the user**
 175 **interfaces (UIs) of LLMs**. This circumvents the requirement to access internal sampling distributions. The method
 176 constructs a keyword set and employs an encrypted steganographic mapping for embedding. It proposes an optimization
 177 mechanism based on reject sampling for accurate extraction and rich semantics [36].
 178

179 Another framework, **Co-Stega**, leverages LLMs to address the challenge of low capacity in social media. It expands
 180 the text space for hiding messages through context retrieval and **increases the generated text's entropy via specific**
 181 **prompts** to enhance embedding capacity. This approach also aims to maintain text quality, fluency, and relevance [19].
 182

183 The concept of **zero-shot linguistic steganography** with LLMs utilizes in-context learning, where samples of
 184 covertext are used as context to generate more intelligible stegotext using a question-answer (QA) paradigm [20]. LLMs
 185 are also employed in approaches like **ALiSa**, which directly conceals token-level secret messages in seemingly natural
 186 steganographic text generated by off-the-shelf BERT [7] models equipped with Gibbs sampling [40].
 187

188 The increasing popularity of deep generative models has made it feasible for provably secure steganography to be
 189 applied in real-world scenarios, as they fulfill requirements for perfect samplers and explicit data distributions (see
 190 Section 2) [10, 16, 25].
 191

192 LLM-based steganographic methods are typically evaluated on two primary axes: imperceptibility (perceptual,
 193 statistical, and cognitive measures) and embedding capacity (e.g., bits per token or bits per word). Imperceptibility
 194 evaluations may include automatic metrics (PPL, Distinct-n, MAUVE, KL/JSD) as well as human judgements; embedding
 195 capacity is usually reported as bits/token or overall embedding rate.
 196

197 We now turn to the principal challenges these models face, including the trade-off between imperceptibility and
 198 capacity, robustness to tokenization, and practical deployment constraints.
 199

200 201 3.3 Challenges and Limitations in Steganography with LLMs

202 3.3.1 *Perceptual vs. Statistical Imperceptibility (Psic Effect)*. The **Psic Effect** [39] represents a fundamental trade-off in
 203 steganographic systems.
 204

205 3.3.2 *Low Embedding Capacity*. Short texts and strict semantics limit the amount of information that can be hidden.
 206

207 [Placeholder footnote]

209 3.3.3 *Lack of Semantic Control and Contextual Consistency.* Ensuring generated text matches intended meaning and
210 context is difficult.
211

212 3.3.4 *Challenges with LLMs in Steganography.* LLMs may introduce unpredictability, bias, or leak information.
213

214 3.3.5 *Segmentation Ambiguity.* Tokenization can cause ambiguity in how information is embedded or extracted.
215

216 A primary challenge in steganography, particularly when utilizing Large Language Models (LLMs), revolves around
217 the distinction between white-box and black-box access (see Section 2.2). Most current advanced generative text
218 steganographic methods operate under a "white-box" paradigm, requiring direct access to the LLM's internal components,
219 such as its training vocabulary and the sampling probabilities of words. This presents a significant limitation because
220 many state-of-the-art LLMs are proprietary and are accessed by users primarily through black-box APIs or user
221 interfaces [36]. Consequently, these white-box methods are often impractical for real-world deployment with popular
222 commercial LLMs. Furthermore, methods that rely on modifying the sampling probability distribution to embed secret
223 messages inherently introduce security risks because they alter the original distribution, making the steganographic
224 text statistically distinguishable from normal text [10, 16, 36, 39].
225

226 Another significant hurdle is **ensuring both the quality and imperceptibility of the generated text**, encompassing perceptual, statistical, and cognitive imperceptibility [8]. While advancements in deep neural networks have
227 improved text fluency and embedding capacity, older models or certain embedding strategies can still produce texts
228 that lack naturalness, logical coherence, or diversity compared to human-written content. Linguistic steganography
229 methods often struggle to control the semantics and contextual characteristics of the generated text, leading to a decline
230 in its "cognitive-imperceptibility" [8, 39]. This can make concealed messages easier for human or machine supervisors
231 to detect. Although models like NMT-Stega and Hi-Stega aim to maintain semantic and contextual consistency by
232 leveraging source texts or social media contexts, this remains a complex challenge [8, 35].
233

234 **Channel entropy requirements and variability** also pose a considerable challenge. Traditional universal steganographic schemes often demand consistent channel entropy, which is rarely maintained in real-world natural language communication. Moments of low or zero entropy can cause protocols to fail or require extraordinarily long steganographic texts. The Psic Effect highlights this dilemma in balancing quality and detectability.
235

236 Furthermore, **segmentation ambiguity** introduced by subword-based language models presents a critical issue for provably secure linguistic steganography. When a sender detokenizes generated subword sequences into continuous text, the receiver might retokenize it differently, leading to decoding errors [25].
237

238 Additional limitations include:
239

- 240 • **Computational Overhead:** LLMs incur 3-5 times higher computational cost than prior methods [20].
241
- 242 • **Data Integrity and Reversibility:** Some methods cannot perfectly recover the original cover text after message
243 extraction [26, 44].
244
- 245 • **Ethical Concerns:** Pre-trained LLMs may introduce biases, discrimination, or inappropriate content [3, 20].
246
- 247 • **Provable Security:** Many NLP steganography works lack rigorous security analyses and fail to meet formal
248 cryptographic definitions [16].
249
- 250 • **Hallucinations:** LLMs can generate factually incorrect or contextually inappropriate content, leading to
251 embedding errors [12].
252
- 253 • **Channel Entropy Limitations:** Short, context-dependent texts have lower entropy, limiting hiding capacity
254 [19].
255

256 [Placeholder footnote]
257

258

259

260

261 4 RELATED REVIEWS

262 5 RESEARCH METHOD

263
264 This study was undertaken as a systematic mapping review using the guidelines presented in Petersen et al. [24]. The
265 goal of this review is to identify, categorize, and analyze existing literature published between 2018 and 2025 and use
266 syntactic and semantics aspects to represent context handling in linguistic steganographic methods.
267

268
269 **5.1 Planning**

270 In this section, we define our research questions, the search strategy we use, and the inclusion and exclusion criteria
271 considered to filter the results.
272

273
274 *5.1.1 Research Questions.* This systematic literature review is guided by six research questions, aiming to comprehen-
275 sively map the landscape of steganographic techniques leveraging large language models (LLMs). The questions explore
276 the current state of published literature, applications where these techniques are being explored, and the metrics and
277 evaluation methods used to assess their performance, with a focus on capacity, security, and contextual compatibility.
278 Furthermore, the review investigates how external knowledge sources are integrated to enhance capacity or contextual
279 relevance, the limitations and trade-offs associated with current techniques, and potential future research directions
280 considering emerging trends and identified gaps.
281

282
283 *5.1.2 Search Strategies.* The initial literature search employed a specific query string: '(steganography or watermark or
284 "Information Hiding") and ("Large Language Model" or LLM or BERT or LAMA or GPT)'. This query was executed
285 across several digital libraries, including ACM Digital Library, IEEE Digital Library, Science@Direct, Scopus, and
286 Springer Link, to ensure broad coverage. To complement this automated search and identify additional relevant studies,
287 a snowballing technique was also applied. This involved examining the reference lists of included studies. While
288 snowballing primarily yielded older steganographic techniques not explicitly mentioning LLMs, these papers often
289 utilized similar methodological approaches to contemporary LLM-based steganography, providing valuable contextual
290 information.
291

292
293 *5.1.3 Inclusion and Exclusion Criteria.* To ensure the selection of high-quality and relevant studies, the following
294 criteria were applied.
295

296 **Inclusion Criteria** Studies were included if they:

- 297** IC1: Provided full-text access.
- 298** IC2: Were published in English from 2018 onwards.
- 299** IC3: Appeared in peer-reviewed journals, conferences, or workshops.
- 300** IC4: Directly addressed steganography, watermarking, or information hiding techniques involving or significantly
301 impacted by LLMs, BERT, LAMA, or GPT architectures.
- 302** IC5: Represented empirical studies, surveys, reviews, or theoretical contributions.

303 **Exclusion Criteria** Studies were excluded if they:

- 304** EC1: Were duplicates (retaining the most complete or recent version).
- 305** EC2: Were incomplete, abstract-only, or irrelevant to steganography with LLMs.
- 306** EC3: Were non-English publications.

307 [Placeholder footnote]

313 EC4: Came from non-peer-reviewed sources (e.g., preprints, dissertations, theses, books, book chapters), unless
314 extended from peer-reviewed conference papers.
315

316 **5.2 Conducting the Search**

317 The initial automated search across the selected digital libraries yielded a total of 1043 candidate papers. The distribution
318 by source was: ACM Digital Library (346), IEEE Digital Library (61), Science@Direct (209), Scopus (151), and Springer
319 Link (276). Duplicated papers were automatically eliminated using Parsifal tool¹. After removing all duplicates, 1,573
320 papers remained. Following this the papers underwent a multi-stage filtering process based on their titles, abstracts, and
321 full texts, guided by the predefined inclusion and exclusion criteria. After title and abstract filtering, 58 papers remained.
322 Of these, 26 were accepted with readily available PDFs, while 6 were pending PDF acquisition at the time of analysis.
323

324 **5.3 Data Extraction and Classification**

325 A Data Extraction Form (DEF) was developed to systematically collect data from each primary study to address our
326 research questions. The form is designed in a table format consisting of the following types of information:
327

- 328 • Bibliometric Information: paper title, type (Steganography or Watermarking), author(s), publication year, and
329 publication venue.
- 330 • Model Details: input and output formats, key characteristics, approach classification (three-term categorical),
331 specific LLM used (if applicable), embedding process description, and code availability.
- 332 • Datasets: all datasets employed, including their sizes.
- 333 • Context Awareness: whether the method is "Explicit," "Implicit," or "No," the context keyword (e.g., "Social
334 Media," "Formal Document"), how context is represented (e.g., "Text," "Pretext," "Graph," "Vector"), and how it is
335 utilized in the method.
- 336 • Evaluation Details: evaluation metrics, steganalysis models used, and the best numerical results for each reported
337 metric.
- 338 • Strengths and Limitations: main strengths and weaknesses of the approach or model.

339 Following data extraction, studies were classified based on predefined categories derived from the research questions
340 to identify trends, patterns, and gaps in the literature. The results are summarized using tables, figures ??), and descriptive
341 statistics. Each research question is addressed individually with interpretation of findings and identification of future
342 research directions.

343 **6 RESULTS**

344 This section presents the synthesized findings from our systematic literature review of 26 primary studies (6 pending
345 acquisition) on LLM-based steganography. The results are organized around five research questions to provide a
346 comprehensive analysis of the current state, applications, evaluation methods, knowledge integration, and limitations
347 in this rapidly evolving field.

348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364

¹<https://parsif.al>

365 6.1 State of Published Literature on LLM-based Steganography (RQ1)

366 Our analysis reveals a significant surge in LLM-based steganography research since 2023, with approximately 17 new
 367 papers published in 2024–2025. The field has evolved from early white-box modifications to more practical hybrid and
 368 black-box approaches.

Year	2020	2021-2022	2023	2024-2025	Total
Publications	2	3	4	17	26

375 Table 1. Publication trends by year

379 6.1.1 Publication Trends and Distribution.

380 6.1.2 Model Preferences and Venues. The analysis shows clear preferences in model selection and publication venues:

- 383 • **Model Usage:** Over 80% of studies utilize open-weight models (GPT-2, LLaMA/LLaMA2, BERT, OPT, BART),
 384 while approximately 12% use proprietary models (GPT-3.5/4, ChatGPT), and 8% employ custom or from-scratch
 385 architectures
- 387 • **Publication Venues:** 60% appear in preprint servers (arXiv), 25% in top-tier conferences (ACL, NeurIPS, ICLR),
 388 and 15% in specialized venues
- 390 • **Geographic Distribution:** 45% from Asia-Pacific, 35% from North America, 20% from Europe

391 6.1.3 Research Gaps and Opportunities. Several significant gaps were identified:

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

398 6.1.4 Key Trends and Evolution. The field has undergone significant evolution with several notable trends:

- 401 • **Paradigm Shift:** Early works (pre-2024) primarily concentrated on white-box modifications, such as token
 402 sampling in GPT-2, whereas recent trends demonstrate a shift toward hybrid and black-box approaches for
 403 more practical, real-world deployment
- 404 • **Model Democratization:** The increasing availability of open-source LLMs has democratized research in this
 405 field
- 407 • **Integration with Watermarking:** Approximately 40% of research integrates concepts from digital watermarking,
 408 creating hybrid approaches
- 409 • **Context Awareness:** Growing emphasis on context-aware steganographic systems that leverage domain-
 410 specific knowledge

412 Recent examples include **FREmax** (2024), which advanced frequency-based sampling for improved imperceptibility,
 413 and **Hi-stega** (2024), which provides a hierarchical retrieval-augmented approach combining high payload with semantic
 414 coherence. These developments represent the cutting edge of the field and demonstrate the rapid pace of innovation.

415 [Placeholder footnote]

417 6.2 Applications of LLM-based Steganographic Techniques (RQ2)

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

423 Application Domain	424 Percentage	425 Studies	426 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

431 Table 2. Distribution of applications across reviewed studies

434 6.2.1 Primary Applications.

436 6.2.2 *Covert Communication Applications.* Covert communication represents the primary application domain, with
 437 approximately 60% of papers focusing on this use case. Key characteristics include:

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

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

447 6.2.3 *Watermarking and Fingerprinting Applications.* About 30% of studies focus on watermarking and fingerprinting
 448 applications:

- 450 • **Content Tracing:** Watermarking for tracking content origin and ownership
- 451 • **Model Fingerprinting:** Identifying and licensing LLMs for commercial use
- 452 • **Copyright Protection:** Embedding ownership information in generated content
- 453 • **Attribution:** Ensuring proper credit for content creators

456 6.2.4 *Emerging Applications.* Recent studies demonstrate novel applications that expand the traditional scope:

- 457 • **Social Media Hiding:** Models such as **Co-Stega** expand text space through context retrieval and entropy
 458 enhancement
- 459 • **Jailbreak Attacks:** Steganography can conceal harmful queries, as demonstrated in **StegoAttack**
- 460 • **Data Exfiltration:** **TrojanStego** embeds secrets directly into LLM outputs
- 461 • **Multimodal Steganography:** Integration with vision-language models for text-image combinations

464 6.2.5 *Domain-Specific Applications.* The field further investigates domain-specific applications, including:

- 466 • **High-Entropy Texts:** Utilization in news articles and formal documents
- 467 • **Short Prompts:** Question-and-answer paradigms for conversational AI

468 [Placeholder footnote]

- **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.

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

[Placeholder footnote]

521 6.3.3 *Capacity Metrics.* Capacity evaluation focuses on embedding efficiency:

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

526 6.3.4 *Security Metrics.* Security evaluation assesses resistance to detection and attacks:

- 527 • **Detection Accuracy:** Performance of steganalysis classifiers
- 528 • **F1 Score:** Balanced precision-recall measure
- 529 • **Attack Resistance:** Performance degradation under various attacks
- 530 • **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

531 Table 5. Performance comparison across method types

532 6.3.5 *Method Comparison.*

533 6.3.6 *Evaluation Methods and Tools.* Evaluation methods encompass both automated tools and human assessment:

- 534 • **Automated Tools:**
 - 535 – Steganalysis classifiers (LS-CNN, BiLSTM-Dense, BERT-FT)
 - 536 – Statistical analysis tools
 - 537 – Semantic similarity measures
- 538 • **Human Evaluation:**
 - 539 – Fluency judgments
 - 540 – Naturalness assessment
 - 541 – Detection difficulty evaluation

542 6.3.7 *Evaluation Challenges and Gaps.* Several significant challenges exist in current evaluation practices:

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

548 6.3.8 *Recent Advances in Evaluation.* Recent studies have introduced more comprehensive evaluation approaches:

- 549 • **Multi-metric Evaluation:** Combining perceptual, statistical, and semantic metrics
- 550 • **Attack-based Testing:** Systematic evaluation against various attack scenarios
- 551 • **Human-AI Collaborative Assessment:** Combining automated and human evaluation

552 [Placeholder footnote]

573 • Cross-domain Evaluation: Testing across different text types and domains

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

577

578 **6.4 Integration of External Knowledge Sources (RQ4)**

579 The integration of external knowledge sources has emerged as a crucial area of research in LLM-based steganography,
580 with 65% of studies incorporating some form of external information. This integration enhances both capacity and
581 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

594 Table 6. External knowledge integration patterns and benefits

595

596 **6.4.1 Knowledge Source Types.**

597 **6.4.2 Semantic Resources Integration.** Semantic resources provide structured knowledge that enhances contextual
598 understanding:

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

606 **607 Co-Stega** demonstrates effective use of semantic resources by leveraging context retrieval and entropy enhancement
608 for social media applications, achieving significant improvements in both capacity and naturalness.

609 **610 6.4.3 Domain Corpora Integration.** Domain-specific corpora provide specialized knowledge for targeted applications:

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

616 **617 FreStega** exemplifies effective corpus integration, using large corpora for distribution alignment and achieving a
618 15% increase in capacity while maintaining imperceptibility.

619 **620 6.4.4 Prompt Engineering and Context Guidance.** Prompt-based approaches leverage external knowledge through
621 strategic prompting:

- 623 • In-context Learning:** Using examples to guide generation

624 [Placeholder footnote]

- 625 • **Few-shot Learning:** Learning from limited examples
- 626 • **Zero-shot Approaches:** No training examples required
- 627 • **Chain-of-thought:** Step-by-step reasoning guidance

629 Zero-shot steganography methods, such as those using LLaMA2-Chat-7B, demonstrate how prompt engineering can
 630 effectively guide steganographic text generation without requiring model fine-tuning.
 631

632 *6.4.5 Integration Benefits and Performance Gains.* External knowledge integration provides several key benefits:
 633

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

639 *6.4.6 Integration Challenges and Trade-offs.* Despite the benefits, knowledge integration introduces several challenges:
 640

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

647 *6.4.7 Integration Strategies and Architectures.* Different integration strategies have been employed:
 648

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

655 Table 7. Knowledge integration strategies and their characteristics
 656
 657
 658

659 *6.4.8 Future Directions in Knowledge Integration.* Several promising directions for future research emerge:
 660

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

666 The integration of external knowledge sources represents a critical advancement in LLM-based steganography,
 667 enabling more sophisticated and context-aware systems. However, the field must address the associated challenges to
 668 realize the full potential of these approaches.
 669

670 **6.5 Limitations and Trade-offs in Current Techniques (RQ5)**

672 Current LLM-based steganographic techniques face several fundamental limitations and trade-offs that constrain their
 673 practical deployment and security guarantees. Understanding these limitations is crucial for advancing the field and
 674 developing more robust solutions.
 675

676 [Placeholder footnote]

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
- **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

[Placeholder footnote]

- 729 • **Capacity Caps:** Tokenization limits maximum achievable capacity
 730
 731

732 **SparSamp** demonstrates these issues, where token ambiguity (TA) reduces accuracy, and **ShiMer** cannot effectively
 733 boost entropy due to tokenization constraints.
 734

735 6.5.6 *Ethical Concerns and Misuse Potential.* The field faces significant ethical challenges that remain largely unad-
 736 dressed:
 737

- 738 • **Bias and Discrimination:** Generated content may perpetuate harmful biases
 739 • **Misuse Potential:** Techniques can be used for malicious purposes
 740 • **Privacy Violations:** Steganographic systems may compromise user privacy
 741 • **Regulatory Compliance:** Lack of frameworks for responsible use
 742

743 **TrojanStego** exemplifies these concerns, as it can embed secrets directly into LLM outputs, potentially enabling
 744 data exfiltration and other malicious activities.
 745

746 6.5.7 *White-box vs. Black-box Trade-offs.* The choice between white-box and black-box approaches involves funda-
 747 mental trade-offs:
 748

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

749 Table 9. Trade-offs between white-box, black-box, and hybrid approaches
 750
 751
 752
 753
 754

755 6.5.8 *Computational and Resource Constraints.* Performance optimization often conflicts with computational efficiency:
 756

- 757 • **Computational Overhead:** Better results typically require more computational resources
 758 • **Memory Requirements:** Large models and external knowledge increase memory needs
 759 • **Real-time Constraints:** Latency requirements may limit optimization options
 760 • **Scalability Issues:** Performance may degrade with increased scale
 761

762 **UTF** demonstrates this trade-off, showing a 5% drop in HellaSwag performance, while **FreStega** requires corpus
 763 access (100 samples) for optimal performance.
 764

765 6.5.9 *Unresolved Challenges and Future Needs.* Several critical challenges remain inadequately addressed:
 766

- 767 • **Provable Security:** Lack of theoretical foundations for security guarantees
 768 • **Robustness:** Limited resilience to advanced attack methods
 769 • **Standardization:** Absence of common evaluation frameworks
 770 • **Ethical Frameworks:** Missing guidelines for responsible development and use
 771 • **Cross-lingual Support:** Poor performance in non-English languages
 772 • **Real-world Deployment:** Limited testing in actual deployment scenarios
 773

774 [Placeholder footnote]
 775
 776
 777
 778
 779
 780

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

6.5.10 *Quantitative Impact Analysis.* Table 10 provides a quantitative overview of the most significant 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.

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]

833 7.3 Addressing the Psic Effect

834 The Perceptual-Statistical Imperceptibility Conflict emerges as the most significant challenge in the field. This funda-
835 mental trade-off between perceptual quality and statistical security affects 80% of studies and results in an average
836 capacity loss of 1-2 bits per word. Future research should focus on:
837

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

843 7.4 The Role of Context and External Knowledge

844 The integration of external knowledge sources has proven crucial for enhancing both capacity and contextual relevance.
845 However, this integration introduces new challenges:
846

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

852 7.5 Ethical Considerations and Responsible Development

853 The review reveals a concerning gap in ethical considerations, with only 10% of studies addressing ethical implications.
854 This represents a significant oversight given the potential for misuse in:
855

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

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

864 7.6 Limitations of the Review

865 Several limitations of this systematic review should be acknowledged:
866

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

874 7.7 Future Research Directions

875 Based on the synthesis of findings, several promising research directions emerge:
876

877 7.7.1 Technical Advancements.

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

882 [Placeholder footnote]
883
884

885 7.7.2 *Methodological Improvements.*

- 886
 887 • **Standardized Evaluation:** Development of common benchmarks and evaluation protocols
 888 • **Human-Centered Design:** Greater emphasis on human evaluation and usability
 889 • **Cross-Language Support:** Extension to non-English languages and cultural contexts
 890 • **Real-World Testing:** Evaluation in actual deployment scenarios

892 7.7.3 *Ethical and Social Considerations.*

- 893
 894 • **Ethical Frameworks:** Development of guidelines for responsible use
 895 • **Bias Mitigation:** Techniques to prevent discrimination and bias propagation
 896 • **Transparency:** Methods for detecting and auditing steganographic content
 897 • **Regulatory Compliance:** Alignment with emerging AI regulations and standards

899 7.8 **Conclusion**

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

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

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

913 8 CONCLUSION

914 This systematic literature review illuminates the profound impact of Large Language Models (LLMs) on linguistic
 915 steganography, demonstrating a clear paradigm shift toward context-aware, generative systems that prioritize imperceptibility,
 916 embedding capacity, and naturalness. Through analysis of 26 primary studies (with 6 pending for full inclusion),
 917 key research questions were addressed, revealing that the published literature is rapidly evolving. Applications now
 918 span secure communication in social media, zero-shot generation, and watermarking overlaps.

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

925 The findings establish that contextual compatibility—leveraging domain correlations and communicative patterns—is
 926 essential for robust steganographic systems. This development paves the way for more sophisticated covert channels
 927 resistant to both human and automated detection. These advancements hold significant implications for information
 928 [Placeholder footnote]

937 security, enabling high-capacity hidden messaging in everyday digital interactions while mitigating risks such as
 938 hallucinations and biases in LLMs.
 939

940 Future research should concentrate on several key areas: mitigating segmentation ambiguity, developing provably
 941 secure black-box frameworks, and exploring multimodal integrations (e.g., text with images) to bridge identified gaps.
 942 This review underscores the potential of LLMs to redefine steganography as a cornerstone of secure, imperceptible
 943 communication in an increasingly surveilled digital landscape.
 944

945 REFERENCES

- 947 [1] 2020. Language Models are Few-Shot Learners. arXiv:[2005.14165](https://arxiv.org/abs/2005.14165) [cs.CL] <https://arxiv.org/abs/2005.14165>
- 948 [2] 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:[2307.09288](https://arxiv.org/abs/2307.09288) [cs.CL] <https://arxiv.org/abs/2307.09288>
- 949 [3] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language
 950 models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. ACM, Virtual Event, Canada, 610–623.
- 951 [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
 952 Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.
- 953 [5] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott
 954 Lundberg, et al. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. *arXiv preprint arXiv:2303.12712* (2023).
- 955 [6] Christian Cachin. 1998. An Information-Theoretic Model for Steganography. In *Information Hiding*, David Aucsmith (Ed.). Springer Berlin Heidelberg,
 Berlin, Heidelberg, 306–318.
- 956 [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language
 957 Understanding. *arXiv preprint arXiv:1810.04805* (2018).
- 958 [8] Changhao Ding, Zhangjie Fu, Zhongliang Yang, Qi Yu, Daqiu Li, and Yongfeng Huang. 2023. Context-aware linguistic steganography model based
 959 on neural machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2023), 868–878.
- 960 [9] Changhao Ding, Zhangjie Fu, Qi Yu, Fan Wang, and Xianyi Chen. 2023. Joint linguistic steganography with BERT masked language model and
 961 graph attention network. *IEEE Transactions on Cognitive and Developmental Systems* 16, 2 (2023), 772–781.
- 962 [10] Jinyang Ding, Kejiang Chen, Yaofei Wang, Na Zhao, Weiming Zhang, and Nenghai Yu. 2023. Discop: Provably secure steganography in practice
 963 based on distribution copies. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Francisco, CA, USA, 2238–2255.
- 964 [11] Jessica Fridrich. 2009. *Steganography in digital media: principles, algorithms, and applications*. Cambridge University Press, Cambridge, UK.
- 965 [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).
- 966 [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](https://doi.org/10.1109/CompComm.2016.7925203)
- 967 [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
 968 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](https://doi.org/10.1121/1.2016299)
- 969 [15] Zhe Ji, Qiansiqi Hu, Yicheng Zheng, Liyao Xiang, and Xinbing Wang. 2024. A principled approach to natural language watermarking. In *Proceedings
 970 of the 32nd ACM International Conference on Multimedia*. ACM, 2908–2916.
- 971 [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.
- 972 [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>
- 973 [18] Fanxiao Li, Sixing Wu, Jiong Yu, Shuxin Wang, BingBing Song, Renyang Liu, Haoseng Lai, and Wei Zhou. 2023. Rewriting-Stego: generating
 974 natural and controllable steganographic text with pre-trained language model. In *International Conference on Database Systems for Advanced
 975 Applications*. Springer, 617–626.
- 976 [19] Guorui Liao, Jinshuai Yang, Kaiyi Pang, and Yongfeng Huang. 2024. Co-stega: Collaborative linguistic steganography for the low capacity challenge
 977 in social media. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*. ACM, Baiona, Spain, 7–12.
- 978 [20] Ke Lin, Yiyang Luo, Zijian Zhang, and Ping Luo. 2024. Zero-shot generative linguistic steganography. *arXiv preprint arXiv:2403.10856* (2024).
- 979 [21] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment
 980 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.
- 981 [22] Mohammed Abdul Majeed, Rossilawati Sulaiman, Zarina Shukur, and Mohammad Kamrul Hasan. 2021. A Review on Text Steganography Techniques.
 982 *Mathematics* 9, 21 (2021). doi:[10.3390/math9212829](https://doi.org/10.3390/math9212829)
- 983 [23] Travis Munyer, Abdullah All Tanvir, Arjon Das, and Xin Zhong. 2024. DeepTextMark: a deep learning-driven text watermarking approach for
 984 identifying large language model generated text. *Ieee Access* 12 (2024), 40508–40520.

985 [Placeholder footnote]

- [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] 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.
- [26] 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.
- [27] 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
- [28] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. *Language Models are Unsupervised Multitask Learners*. Technical Report. OpenAI.
- [29] 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
- [30] Murray Shanahan. 2024. Talking about large language models. *Commun. ACM* 67, 2 (2024), 68–79.
- [31] Gustavus J Simmons. 1984. The prisoners' problem and the subliminal channel. In *Advances in Cryptology: Proceedings of Crypto 83*. Springer, Boston, MA, 51–67.
- [32] Martin Steinebach. 2024. Natural language steganography by chatgpt. In *Proceedings of the 19th International Conference on Availability, Reliability and Security*. ACM, 1–9.
- [33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).
- [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>
- [38] Aiyuan Yang, Bin Xiao, Binyuan Wang, Binxin Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open Large-scale Language Models. *arXiv preprint arXiv:2309.10305* (2023).
- [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]

1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081

Table 11. Summary of Results from Reviewed Papers

Paper	Llm	Year	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	2020.0	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	2022.0	BookCorpus	BPW=0.5335 F1=0.9402 PPL=134.2199	non-explicit	pre-text	text
Co-stega: Collaborative linguistic steganograph... [19]	Llama-2-7B-chat, GPT-2 (fine-tuned), Llama-2-13B	2024.0	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

Paper	Llm	Year	Dataset	Result	Context Aware	Categ Context	Representation Context
Joint linguistic steganography with BERT masked... [9]	LSTM + attention for temporal context. GAT for spatial token relationships. BERT MLM for deep semantic context in substitution.	2023.0	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	2023.0	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 + pretext	text

Continued on next page

[Placeholder footnote]

Table 11 – continued from previous page

Paper	Llm	Year	Dataset	Result	Context Aware	Categ Context	Representation Context
Generative text steganography with large language... [36]	Any	2024.0	[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 ... [16]	GPT-2	2021.0	Hutter Prize, HTTP GET requests	GPT-2: 3.09 bits/token	non-explicit	tuning + pretext	text
Zero-shot generative linguistic steganography [20]	LLaMA2-Chat-7B (as the stegotext generator / QA model). GPT-2 (for NLS baseline and JSD evaluation)	2024.0	IMDB, Twitter	PPL: 8.81. JSDfull: 17.90 (x10[truncated]iiciircum-2). JSDhalf: 16.86 (x10[truncated]iiciircum-2). JSDzero: 13.40 (x10[truncated]iiciircum-2) TS...	explicit	zero-shot prompt	+ text

Continued on next page

[Placeholder footnote]

Paper	Llm	Year	Dataset	Result	Context Aware	Categ Context	Representation Context
Provably secure disambiguating neural linguisti... [25]	LLaMA2-7b (English), Baichuan2-7b (Chinese)	2024.0	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	2024.0	Web Transformer 2	Bit acc: 0.994 (K=None), 1.000 (DAE), 0.978 (Adaptive+K=S); Meteor Drop: [truncated]iitilde0.057; SBERT ↑: [truncated]iitilde1.227; Ownership R...	Yes; semantic-level embedding; synonym substitution using BERT	Yes; watermark message assigned categorical label (e.g., 4-bit → 1-of-16)	Yes; semantic embeddings via transformer encoder and BERT; SBERT distance as metric

Continued on next page

Table 11 – continued from previous page

Paper	Llm	Year	Dataset	Result	Context Aware	Categ Context	Representation Context
Context-aware linguistic steganography model ba... [8]	BERT (encoder), LSTM (decoder)	2024.0	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] untilde16%	Yes	[Not specified]	GCF (global context), LMR (language model reference), Multi-head attention
DeepTextMark: a deep learning-driven text water... [23]	Model-independent; tested with OPT-2.7B	2024.0	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... [35]	GPT-2	2024.0	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

Paper	Llm	Year	Dataset	Result	Context Aware	Categ Context	Representation Context
Linguistic steganography: From symbolic space t... [41]	CTRL (generation), BERT (semantic classifier)	2020.0	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]iiitilde0.5	implicit	Text	Semanteme (α) as a vector in semantic spac
Natural language steganography by chatgpt [32]	[Not specified]	2024.0	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... [26]	Transformer (Paraphraser), BART (BARTScore), BERT (BLEURT, comparisons)	2023.0	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]iiitilde88–90%	Explicit	[Not specified]	text

Continued on next page

[Placeholder footnote]

Table 11 – continued from previous page

Paper	Llm	Year	Dataset	Result	Context Aware	Categ Context	Representation Context
Rewriting-Stego: generating natural and control... [18]	BART (bart-base2)	2023.0	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)	2022.0	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]
Imperceptible Text Steganography based on Group...	Qwen-7B-Chat	2024.0	HC3, DailyDialogue, COCO Descriptions	HC3: Bit 188.94, Stego 131.99, PPL 34.07, Mean 20.19, Var 0.1e04, F1 90.01%; DailyDialogue: Bit 188.94, Stego 89.37, PPL 53.88, Mean 20.13, Var 0....	Explicit	Social Media / Group Chat	Text (chat history and current input)

Continued on next page

Paper	Llm	Year	Dataset	Result	Context Aware	Categ Context	Representation Context
A Semantic Controllable Long Text Steganography...	Llama 7B Chat, Meta LLaMA2 7B Chat	2024.0	Story (ChatGPT), Post (Recipe Kaggle + ChatGPT), Ad (Mobile Kaggle + ChatGPT)	ppl ↓ >23%, Δppl ↓ >72% vs ADG/HC/Bin; detection accuracy ↓ >10% vs baselines	Explicit	Topical Content	KG triplets (e1, r, e2), task descriptions (D)
Beyond Binary Classification: Customizable Text...	gpt-3.5-turbo-instruct, OPT-6.7b, babbage-002, davinci-002 (others: ChatGPT, GPT-2-4, LLaMA)	2024.0	Realnewslike (C4, 500 samples, 100-token prompts + completions); Custom watermark dataset (short info <10 tokens)	AUC 0.98, FPR 0.00, FNR 0.00, [truncated]single-letter decoding, PPL close to human text	Implicit	General Text Generation	Text (evolving prompt + generated output)
CPG-LS: Causal Perception Guided Linguistic Ste...	BERTBase, Cased	2023.0	CC-100 corpus; 10k cover texts; 7:3 train-test split	PPL 36.5; Mauve 0.871; Payload 0.150 bits/word; BiLSTM-D Acc 0.387 F1 0.375; R-BI-C Acc 0.378 F1 0.366; TS-RNN Acc 0.380 F1 0.368	Implicit	Natural Language Text	Text, embeddings, vector matrix

Continued on next page

Table 11 – continued from previous page

Paper	Llm	Year	Dataset	Result	Context Aware	Categ Context	Representation Context
Controllable Semantic Linguistic Steganography ...	BERT + CRF	2024.0	Gigaword; CNN/Daily Mail	Rouge-1: 0.2212; Rouge-2: 0.0268; Rouge-L: 0.1609; Meteor: 0.1384; Cosine: 0.5911; Euclidean: 5.6386; Manhattan: 87.9534; Jaccard: 0.2022; Anti-ste...	Explicit	Social Media	Semantic features of input text; 384-dim dense vectors for evaluation
FREmax: A Simple Method Towards Truly Secure Ge...	GPT-2	2024.0	Tweet corpus (2.6M sents, 26.8M tokens), IMDB corpus (1.05M sents, 25.3M tokens)	Tweet: PPL 361.83, Entropy 48.21, Tokens 10.83, Distinct3 0.98, BPS 62.79, SI% 73.03. IMDB: PPL 169.66, Entropy 103.39, Tokens 23.80, Distinct3 0....	Implicit	General Text	N-gram frequency distribution stored in a look-up table

[Placeholder footnote]

	CONTENTS	
1410		
1411		
1412	Abstract	1
1413	1 Introduction	1
1414	2 Background	3
1415	3 Steganography and Large Language Models	3
1416	3.1 Capabilities and Approximating Natural Communication	3
1417	3.2 Role in Generative Linguistic Steganography	4
1418	3.3 Challenges and Limitations in Steganography with LLMs	4
1419	3.3.1 Perceptual vs. Statistical Imperceptibility (Psi Effect)	4
1420	3.3.2 Low Embedding Capacity	4
1421	3.3.3 Lack of Semantic Control and Contextual Consistency	5
1422	3.3.4 Challenges with LLMs in Steganography	5
1423	3.3.5 Segmentation Ambiguity	5
1424	4 Related Reviews	6
1425	5 Research Method	6
1426	5.1 Planning	6
1427	5.1.1 Research Questions	6
1428	5.1.2 Search Strategies	6
1429	5.1.3 Inclusion and Exclusion Criteria	6
1430	5.2 Conducting the Search	7
1431	5.3 Data Extraction and Classification	7
1432	6 Results	7
1433	6.1 State of Published Literature on LLM-based Steganography (RQ1)	8
1434	6.1.1 Publication Trends and Distribution	8
1435	6.1.2 Model Preferences and Venues	8
1436	6.1.3 Research Gaps and Opportunities	8
1437	6.1.4 Key Trends and Evolution	8
1438	6.2 Applications of LLM-based Steganographic Techniques (RQ2)	9
1439	6.2.1 Primary Applications	9
1440	6.2.2 Covert Communication Applications	9
1441	6.2.3 Watermarking and Fingerprinting Applications	9
1442	6.2.4 Emerging Applications	9
1443	6.2.5 Domain-Specific Applications	9
1444	6.2.6 Application Requirements and Constraints	10
1445	6.3 Evaluation Metrics and Methods (RQ3)	10
1446	6.3.1 Metric Categories and Standards	10
1447	6.3.2 Imperceptibility Metrics	10
1448	6.3.3 Capacity Metrics	11
1449	6.3.4 Security Metrics	11
1450	6.3.5 Method Comparison	11
1451		
1452	[Placeholder footnote]	

1462	6.3.6	Evaluation Methods and Tools	11
1463	6.3.7	Evaluation Challenges and Gaps	11
1464	6.3.8	Recent Advances in Evaluation	11
1465	6.4	Integration of External Knowledge Sources (RQ4)	12
1466	6.4.1	Knowledge Source Types	12
1467	6.4.2	Semantic Resources Integration	12
1468	6.4.3	Domain Corpora Integration	12
1469	6.4.4	Prompt Engineering and Context Guidance	12
1470	6.4.5	Integration Benefits and Performance Gains	13
1471	6.4.6	Integration Challenges and Trade-offs	13
1472	6.4.7	Integration Strategies and Architectures	13
1473	6.4.8	Future Directions in Knowledge Integration	13
1474	6.5	Limitations and Trade-offs in Current Techniques (RQ5)	13
1475	6.5.1	Key Limitations	14
1476	6.5.2	The Psic Effect: A Fundamental Trade-off	14
1477	6.5.3	Attack Vulnerability and Security Concerns	14
1478	6.5.4	Capacity Limitations in Short Texts	14
1479	6.5.5	Segmentation and Tokenization Issues	14
1480	6.5.6	Ethical Concerns and Misuse Potential	15
1481	6.5.7	White-box vs. Black-box Trade-offs	15
1482	6.5.8	Computational and Resource Constraints	15
1483	6.5.9	Unresolved Challenges and Future Needs	15
1484	6.5.10	Quantitative Impact Analysis	16
1485	7	Discussion	16
1486	7.1	Synthesis of Key Findings	16
1487	7.2	Implications for Research and Practice	16
1488	7.2.1	Methodological Implications	16
1489	7.2.2	Practical Implications	16
1490	7.3	Addressing the Psic Effect	17
1491	7.4	The Role of Context and External Knowledge	17
1492	7.5	Ethical Considerations and Responsible Development	17
1493	7.6	Limitations of the Review	17
1494	7.7	Future Research Directions	17
1495	7.7.1	Technical Advancements	17
1496	7.7.2	Methodological Improvements	18
1497	7.7.3	Ethical and Social Considerations	18
1498	7.8	Conclusion	18
1499	8	Conclusion	18
1500	References		19

	Contents	30
1514		
1515		
1516		
1517		
1518		
1519		
1520		
1521		
1522		
1523		
1524		
1525		
1526		
1527		
1528		
1529		
1530		
1531		
1532		
1533		
1534		
1535		
1536		
1537		
1538		
1539		
1540		
1541		
1542		
1543		
1544		
1545		
1546		
1547		
1548		
1549		
1550		
1551		
1552		
1553		
1554		
1555		
1556		
1557		
1558		
1559		
1560		
1561		
1562		
1563		
1564		
1565	[Placeholder footnote]	