LLM ステガノグラフィの流暢性向上に向けた基礎的検討

井戸 智斗志[1,a)]　石田 繁巳[2]　稲村 浩[2]

概要：本稿では，大規模言語モデル（LLM）を用いたテキストステガノグラフィにおいて，生成文章の流暢性を高める手法を提案する．既存手法では秘密文章の埋め込み時に低確率なトークンを選択せざるを得ず，文脈の不自然さや文末の唐突な途切れが生じるという課題があった．これに対し，提案手法では，トークンの確率順位を引き上げる写像を行った上で埋め込む手法に加え，文末までの生成制御およびカバー文章の未完結化を導入することで生成文章の流暢性を向上させる．評価実験では，Perplexity（PPL）を指標とした流暢性と LanguageTool による文法評価を行い，従来手法と比較して PPL を 1/2.8 倍に，エラー率を 1/3.3 倍に低下できることを示した．さらに，規模や構造が異なる 3 つモデルを用いた検証により，提案手法が特定のモデルに依存していないことを確認した．

1　はじめに

ステガノグラフィは，画像 [1] や音声 [2]，テキスト [3] などの媒体の中に，秘密情報を第三者に気づかれないように埋め込む技術である．秘密情報が埋め込まれた成果物は「ステゴデータ」と呼ばれ [4]，情報秘匿の重要な手段として研究されてきた．従来の画像や音声を用いた手法は，データのバイナリ的な冗長性を利用するものが主流であったが，これらは媒体のデータサイズやフォーマットによる制約を強く受ける．これに対し，テキストを媒体とするテキストステガノグラフィは，データ容量が小さく，かつ通信プロトコルを選ばず SNS やメール，文書ファイルなどあらゆる場面で利用できるため，高い汎用性を有している．

こうした背景のもと，大規模言語モデル（LLM）の飛躍的な発展が，テキストステガノグラフィの可能性をさらに広げつつある．LLM は統計的に自然な文章を生成する能力に長けており，それまでに生成されたトークンの文脈から，次のトークンの確率分布を計算し，高確率のトークンを選択することで自然な文章を生成している．この仕組みを利用して，秘密文章を主語・述語・目的語・感情という 4 つの要素に変換し，プロンプトを用いて LLM にそれらを含む自然な文章を生成させることでステゴ文章を生成する手法 [3] や，秘密文章の埋め込み先となる「カバー文章」に対して秘密文章のトークン確率遷移に基づいて「ステゴ文章」を生成する手法 [5] などが提案されている．

しかしながら，従来の LLM ステガノグラフィ手法にはステゴ文章の流暢性が損なわれるという課題がある．本稿における流暢性とは，生成された文章が文法的に正しく，人間が読んで自然であるという文章の品質を指す．従来手法では秘密文章を埋め込むために LLM にとって確率が低いトークンを選択する必要性が生じる場合があり，単語間の文脈的な連続性が失われ，流暢性が損なわれる．

これに対し，本研究では，任意の秘密文章から流暢なステゴ文章を生成する LLM ステガノグラフィ手法を提案する．本手法のキーアイデアは，ステゴ文章生成時に選択されるトークンの確率順位を引き上げることである．文献 [5] と同様に，「カバー文章」に対して秘密文章のトークン確率遷移に基づいて文章を生成することで秘密文章を埋め込んだステゴ文章を生成するが，その際に選択されるトークンの確率順位を引き上げ，LLM にとって確率の高いトークンを選択させる．その結果，LLM にとって予測しやすい自然な文章がされ，流暢性が向上する．トークンの確率順位を引き上げるアプローチの一例として，本稿では，秘密文章から得られるトークンの確率順位を 2 進数で表現して「ビット分割」してから埋め込む手法を示す．

ステゴ文章の流暢性をさらに高めるため，本稿では，秘密文章の埋め込み先となる「カバー文章」の未完結化と，ステゴ文章の終端制御という 2 つの補助手法を導入する．カバー文章の未完結化によって，カバー文章と秘密文章の接続部における文脈的不自然さを低減する．そして，ステゴ文章の終端制御として，秘密文章の埋め込み終了後も適切な確率順位の範囲からトークンを選択して文末記号まで生成を継続し，ステゴ文章が文末に達しない問題を抑制する．

提案手法がステゴ文章の流暢性向上に寄与することを検証するために，言語モデルの予測確信度を示す指標である

[1] 公立はこだて未来大学大学院 システム情報科学研究科
[2] 公立はこだて未来大学 システム情報科学部
[a)] g2125004@fun.ac.jp

Perplexity（PPL）[6] と，校正エンジン LanguageTool [7] を用いて文章の流暢度を評価した．その結果, PPL は 1/2.8 倍に，LanguageTool による評価では誤り率を 1/3.3 倍に低下できることを確認した．さらに，提案手法の汎用性を検証するために，異なる LLM を用いて同様の評価を行ったところ，全モデルにおいて PPL で同様の改善傾向が観察され，提案手法が特定のモデルに依存せず広く適用可能である可能性が示唆された．

本稿の構成は以下の通りである．第 2 節では LLM を用いたステガノグラフィに関する関連研究を示す．第 3 節では LLM ステガノグラフィの具体的な手順を示す．第 4 節では提案する流暢性向上の手法を示し，第 5 節において PPL と LanguageTool を用いて，提案手法の流暢性を評価する．最後に第 6 節でまとめとする．

2　関連研究

LLM の台頭は従来のテキスト改変型手法から，トークンの確率分布を直接制御する「生成型言語ステガノグラフィ」への転換をもたらした [8]．本節では，LLM を用いたステガノグラフィを，モデル内部パラメータや確率分布へのアクセス可否に基づいてブラックボックス型とホワイトボックス型に分類し，それぞれの特徴と実運用上の課題について述べる．

ブラックボックス型は, 商用 LLM の API やウェブ UI のみを介して秘匿を行う手法である．Wu らは秘密情報を主語・述語・目的語・感情という 4 つのキーワードに変換し，プロンプトを用いて LLM にそれらを含む自然な文章を生成させることでステゴ文章を生成する手法をした [3]．プロンプトエンジニアリングとキーワードマッピングを組み合わせ，棄却サンプリングによるフィードバック最適化を導入することで，モデル内部に介入せずとも高い復元精度を実現している．さらに，マルチモーダル LLM（MLLM）を活用し，公開画像と説明文章からセッションごとに動的辞書を生成することで，実用性とセキュリティを向上させるブラックボックス型のテキストステガノグラフィも提案されている [9]．固定された辞書を事前に共有する代わりに，公開画像から得られるシード単語と公開辞書を基にして，送受信者が独立して同一の辞書をオンデマンドで構築する．

しかし，ブラックボックス型はモデルの出力のみに依存するため，埋め込める情報が秘密文章の特定のキーワードに限定されやすく，任意の秘密文章を完全に表現することが困難である．そのため，補助媒体として画像が必要とするケースもあり，テキストのみによる自由な埋め込みには制限がある．

ホワイトボックス型は，LLM の内部パラメータや推論時の確率分布を直接操作して秘匿を行う手法であり，ブラックボックス型における「任意の文章を埋め込めない」という制限を克服できるアプローチである．Norelli らは，LLM がトークンの確率遷移モデルとして文章を表現する仕組み [8] に着目し，秘密文章の確率遷移に基づいて別の文章から LLM を用いてステゴ文章を生成する手法を示している [5]．秘密文章の各トークンが持つ確率順位を取得し，別の文章生成時の選択指標として同期させることで，情報の完全な隠蔽と正確な復元を実現している．最大の特徴は両文章が等長となる点であり，ステゴ文章の統計的妥当性は人間が書いた文章の分布内に収まることが実証されている．さらに，受信者の正確な復元を妨げるトークンがないかを送信前に予め検証する手法も提案されている [10]．

しかし，ホワイトボックス型にも実運用上の重大な課題が存在する．ホワイトボックス型はテキストのみで任意の秘密文章を埋め込むことができ，かつ生成されたステゴ文章が統計的に妥当であるという強みを持つ反面，秘密トークンの確率順位が低い場合には，生成モデル側で低確率なトークンを強制的に選択せざるを得ない．その結果，文脈のつながりや文法的な流暢性が損なわれ，人間が読んだ際に違和感を持つ文章が生成される可能性が高くなる．ステガノグラフィの実運用において，第三者に秘密情報の存在そのものを察知されないためには，検知器をすり抜ける統計的な妥当性だけでなく，人間が目にした際の自然さや流暢性を担保することが不可欠である．

したがって，実運用を想定した場合，任意の秘密文章をテキストのみで埋め込み可能というホワイトボックス型の利点を活かしつつ，生成文の流暢性と人間読みにおける自然さを両立させる手法が必要である．そこで本研究では，ホワイトボックス型 LLM ステガノグラフィをベースとして，人間に対する秘匿性を高めるために，トークン選択プロセスにおける流暢性を改良する手法を提案する．

3　LLM ステガノグラフィ

本節では，提案手法のベースとなる，文献 [5] で示されたホワイトボックス型の LLM ステガノグラフィの手順を示す．この手法は LLM によってトークン・確率順位の変換・逆変換を行うことによって実現される．まず，LLM によるトークン・確率順位の変換に関する記法を定義した上で，ステゴ文章の生成プロセス及び秘密文章の復元プロセスを示す．

3.1　LLM によるトークン・確率順位変換に関する記法

LLM を $\mathcal{M}$ とし, 文章は LLM $\mathcal{M}$ を用いてトークン列に変換した状態で表現するものとする．ある文章 $X = (x_1, \ldots)$ の接頭辞 $x_{<i}$ に対する次トークン確率分布，すなわち，LLM にトークン $(x_1, \ldots, x_{i-1})$ を入力した状態での次トークンの確率分布を $P_{\mathcal{M}}(x_{<i})$ とする．文章 X から得られる確率順位系列を $R^X = \left(r_1^X, \ldots\right)$ とすると，

$$r_i^X = \mathrm{rank}_{\downarrow}(P_{\mathcal{M}}(x_{<i}),\, x_i) \tag{1}$$

となる．ここで，$\mathrm{rank}_{\downarrow}(P, \xi)$ は確率分布 P におけるトークン ξ の確率降順での順位を表している（順位の最上位は 0 とする）．

LLM に確率順位系列を与えることで文章，すなわちトークンの系列を生成できる．文章 $X = (x_1, \ldots)$ から得られた確率順位系列 $R^X = \left(r_1^X, \ldots\right)$ を用いて文章 $Y = (y_1, \ldots)$ を生成する場合を考える．確率順位に基づいた LLM によるトークン選択により，生成される文章のトークン y_i は

$$y_i = \mathrm{token}\left(P_{\mathcal{M}}(y_{<i}),\, r_i^X\right) \tag{2}$$

となる．ここで，$\mathrm{token}(P, \rho)$ は確率分布 P において順位が ρ 番目のトークンを表している．文章 X から確率順位系列 R^X を得たときと同一の LLM を用いて初期状態からトークンを選択していけば，確率分布 $P_{\mathcal{M}}(y_{<i})$ の遷移は確率順位系列 R^X を得たときと等しくなり，$y_i = x_i$ となる．

3.2 生成プロセス

送信者は，カバー文章 $C = (c_1, \ldots, c_M)$ と秘密文章 $S = (s_1, \ldots, s_N)$ を入力とし，ステゴ文章 $T = (t_1, \ldots, t_L)$ を生成する．

図 1 に，ステゴ文章生成プロセスの概要を示す．ステゴ文章の生成は 3 つのステップで構成される．

ステップ 1 では，カバー文章 C と秘密文章 S のそれぞれから式 (1) の $\mathrm{rank}_{\downarrow}(P, \xi)$ を用いて確率順位系列 R^C, R^S を得る．

$$R^C = \left(r_1^C, \ldots, r_M^C\right), \quad r_i^C = \mathrm{rank}_{\downarrow}(P_{\mathcal{M}}(c_{<i}), c_i) \tag{3}$$

$$R^S = \left(r_1^S, \ldots, r_N^S\right), \quad r_i^S = \mathrm{rank}_{\downarrow}(P_{\mathcal{M}}(s_{<i}), s_i) \tag{4}$$

ステップ 2 では，確率順位系列 R^C, R^S を連結して埋め込み対象の順位系列 $R^* = (r_1^*, \ldots)$ を得る．

$$R^* = R^C \| R^S = \left(r_1^C, \ldots, r_M^C, r_1^S, \ldots, r_N^S\right) \tag{5}$$

ここで $\|$ は系列連結を表している．

ステップ 3 では，順位系列 R^* から式 (2) の $\mathrm{token}(P, \rho)$ を用いてトークンを選択し，ステゴ文章 T を得る．確率順位系列 R^* の長さは $|R^*| = \left|R^C\right| + \left|R^S\right| = M + N$ であるから，トークン選択を $(M + N)$ 回繰り返すことでステゴ文章 T が得られる．

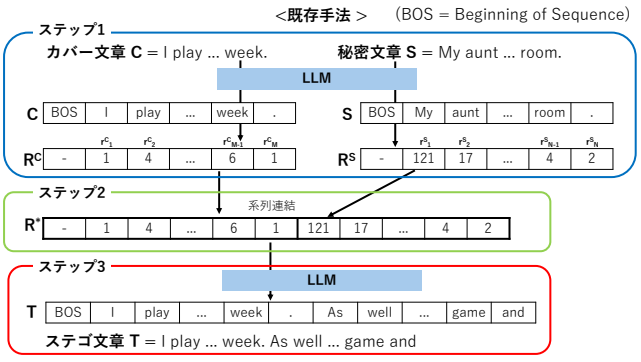

図 1 ステゴ文章生成プロセス

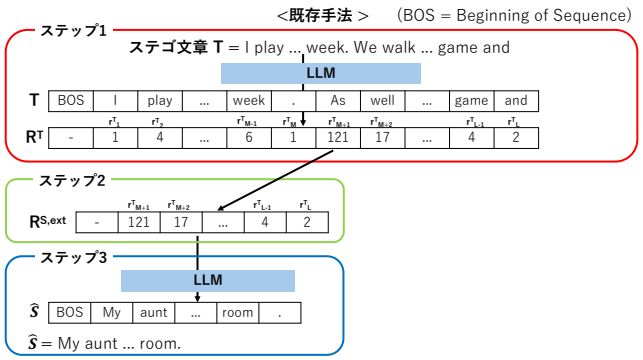

図 2 秘密文章復元プロセス

$$T = (t_1, \ldots, t_{M+N}), \quad t_i = \mathrm{token}\left(P_{\mathcal{M}}(t_{<i}), r_i^*\right) \tag{6}$$

送信者は，得られたステゴ文章 T に加えて，共有情報としてカバー長 M，使用 LLM を受信者に伝達する．

3.3 復元プロセス

図 2 に，秘密文章復元プロセスの概要を示す．秘密文章の復元は 3 つのステップで構成される．

ステップ 1 では，ステゴ文章 T から式 (1) の $\mathrm{rank}_{\downarrow}(P, \xi)$ を用いて確率順位系列 R^T を得る．

$$R^T = \left(r_1^T, \ldots, r_L^T\right), \quad r_i^T = \mathrm{rank}_{\downarrow}(P_{\mathcal{M}}(t_{<i}), t_i) \tag{7}$$

ステップ 2 では，共有されているカバー長 M に基づいて R^T の部分系列を得ることで，秘密文章の順位系列 $R^{S,\mathrm{ext}}$ を取り出す．

$$R^{S,\mathrm{ext}} = \left(r_{M+1}^T, \ldots, r_L^T\right) \tag{8}$$

3.2 節で示した生成プロセスで生成されるステゴ文章の長さは $L = M + N$ であるから，取り出した秘密文章の順位系列の長さは $\left|R^{S,\mathrm{ext}}\right| = N$ である．

ステップ 3 では，順位系列 $R^{S,\mathrm{ext}}$ から式 (2) の $\mathrm{token}(P, \rho)$ を用いてトークンを選択し，秘密文章 $\hat{S}$ を得る．

$$\hat{S} = (\hat{s}_1, \ldots), \quad \hat{s}_i = \mathrm{token}\left(P_{\mathcal{M}}(\hat{s}_{<i}), r_{M+i}^T\right) \tag{9}$$

ステゴ文章生成時と同一の LLM を用いていれば $R^T = R^*$ であるから，式 (4),(5) より $R^{S,\mathrm{ext}} = \left(r_1^S, \ldots, r_N^S\right) = R^S$ となるため，$\hat{S} = S$ である．

4 流暢性向上手法

4.1 キーアイデア

本手法のキーアイデアは，ステゴ文章生成時に選択されるトークンの確率順位を引き上げることで，流暢性を向上させることである．順位系列 R^* の各要素である順位を全

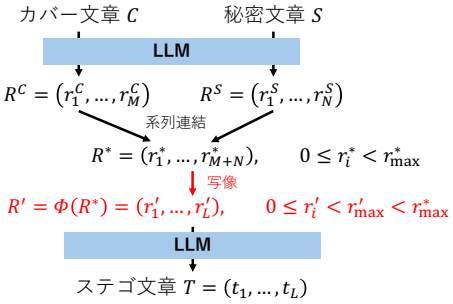

図 3　キーアイデアの概要

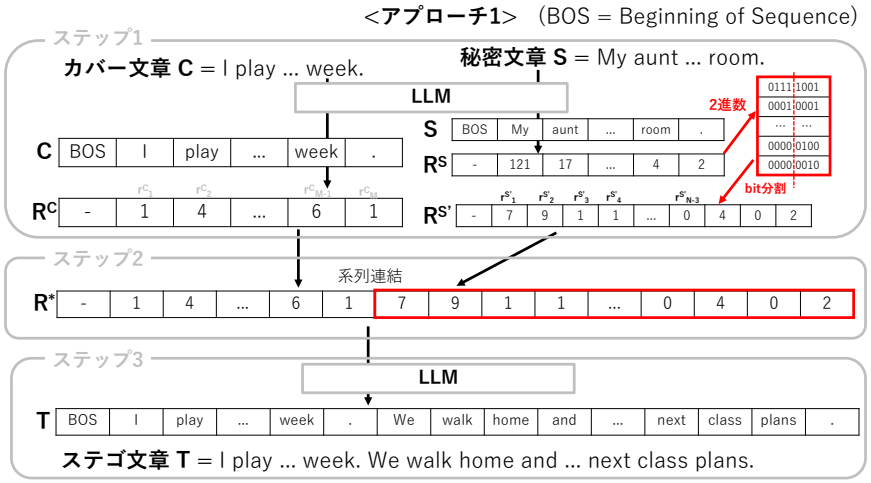

図 4　確率順位のビット分割の例（ビット長 $B = 8$，分割ビット数 $k = 4$）

体的に引き上げる，すなわち順位の値を小さくすることで高確率なトークンを選択し，LLM にとって自然な文章を生成させる．

図 3 に，キーアイデアの概要を示す．3.2 節で示した生成プロセスで得られるステゴ文章 T の確率順位系列 $R^* = (r_1^*, \ldots)$ は $(M+N)$ 次元のベクトルである．確率順位系列が自然な文章から作成されることから確率順位系列 R^* の要素 r_i^* には上界が存在し，$0 \le r_i^* < r_{\max}^*$ であると仮定できる．この上界を小さくするため，ベクトル R^* を L 次元のベクトル $R' = (r_1', \ldots, r_L')$ に写像する．

$$R' = \Phi(R^*) \tag{10}$$

このとき，適切な写像関数 Φ を用いることで，R' の要素 r_i' の上界 $r_{\max}'$ が $r_{\max}^*$ よりも小さくなるようにする．なお，写像関数 Φ にはその逆関数 Φ^{-1} が存在して R' から R^* が一意に求まるものとし，Φ 及び Φ^{-1} を共有情報として受信者に伝達する．

本提案では写像関数 Φ を限定しない．一意な逆変換が存在することが前提となるため，Φ の写像先（ベクトルの総パターン数）が入力となる R^* の総パターン数以上であることが必須となり，

$$(r_{\max}^*)^{M+N} \le (r_{\max}')^L \tag{11}$$

を満たす必要がある．このため，$L > M+N$ となる．

本稿では写像関数 Φ の一例として，R^* の各要素を 2 進数で表してビット分割する手法を示す．

4.2　ビット分割による流暢性向上手法の概要

ビット分割による流暢性向上手法は，確率順位のビット分割，ステゴ文章の文末生成制御，カバー文章の未完結化の 3 つのアプローチによって構成される．以降では，各アプローチについて詳述する．

4.3　アプローチ 1: 確率順位のビット分割

確率順位のビット分割では，写像関数 Φ として，確率順位系列 R^* の各要素を 2 進数で表した上で固定ビット幅に分割する．ビット長 B を $2^B \ge r_{\max}^*$ となるように定め，R^* の各要素を B ビットの 2 進数で表現し，k ビットごとに分割する．ただし，$k < B$ とする．R^* の各要素が B ビットとならない場合には，上位ビットを 0 で埋めた上で分割する．R^* の各要素が分割されることから，分割によって得られる確率順位系列 R' の長さは

$$L = (M+N)\left\lceil \frac{B}{k} \right\rceil \tag{12}$$

となる．

実用上は，カバー文章から得られる確率順位系列 R^C の確率順位の値が小さい場合にはビット分割を行うのは秘密文章から得られる確率順位系列 R^S のみを対象として良い．この場合，

$$L = M + N\left\lceil \frac{B}{k} \right\rceil \tag{13}$$

となる．

図 4 は，$B = 8$，分割ビット数 $k = 4$ の場合における確率順位のビット分割の例を示している．$r_1^S = 121$ であるから，これを 2 進数で表現すると

$$(01111001)_2$$

となり，$k = 4$ ビットごとに分割すると $(0111)_2$ と $(1001)_2$，すなわち，7 と 9 に分割される．これらを新たな $r_1^{S'}, r_2^{S'}$ などとして確率順位系列 $R^{S'}$ を作成し，R^C と連結してステゴ文章を生成する．

4.4　アプローチ 2: ステゴ文章の文末生成制御

ステゴ文章の生成は確率順位系列 R' の長さで打ち切られるため，ステゴ文章が途中で切れたものとなる場合がある．途切れた文章は著しく流暢性を欠いたものであり，第

三者から違和感を持たれる．

そこで，アプローチ 2 として，ステゴ文章の生成を文末記号が選択されるまで継続する．具体的には，確率順位系列 R' からのステゴ文章生成において，確率順位系列 R' の終端に到達した後もランダムなトークンを選択し，文末記号（「.」「!」「?」のいずれか）が得られるまで生成を継続する．このとき，一定値以下の確率順位のトークンをランダムに選択することで流暢性を保つ．

このアプローチでは，ステゴ文章から秘密文章を復元する際に秘密文章のトークン長 $Q = N\lceil B/k \rceil$ が必要となることから，カバー文章と秘密文章の間に秘密文章のトークン長を挿入する．Q が確率順位範囲を超える場合，すなわち $Q \geq 2^B$ の場合にはアプローチ 1 と同様にビット分割する．

4.5 アプローチ 3: カバー文章の未完結化

LLM を用いて文章を生成する際，文章の先頭トークンは，文章の途中にあるトークンに比べて出現確率が低下し，確率順位の数値が大きくなる傾向がある．直前のカバー文章がピリオド等で完結することにより，後続する新たな文の展開における自由度が高くなり，選択され得るトークンの候補が急激に広がるためである．その結果，カバー文章と秘密文章の確率順位系列を連結して得られる R^* や R' は，両者の接続境界において確率順位の数値が突発的に増大する．LLM 本来の予測から外れた低確率なトークンを選択すると，接続境界における文章の流暢性が著しく損なわれる．

そこで，アプローチ 3 として，カバー文章の末尾をピリオド等で完結させず，接続詞や未完結な句構造で終端させることで，後続する秘密文章のトークンとの文脈的一貫性を維持する．“However,” や “As soon as” などの接続句で終了する「未完結状態」の文章をカバー文章として用いることで，LLM は直前の接続句の文脈を引き継いだ状態で後続のトークンを予測するため，確率順位の急激な低下を抑制し，接続境界の流暢性を向上させる．

4.6 ステゴ文章の例

表 1 に，文献 [5] の既存手法及び提案手法によって生成されたステゴ文章の例を示す．既存手法では秘密文章から生成された部分の文章が途中で終わっていたり，「’」が通常は存在しない位置に存在するなど，不自然な点が見られる．これに対し，提案手法では文章が文末まで生成され，一貫した内容を保持していることが分かる．

5 評価

提案手法がステゴ文章の流暢性向上に寄与することを検証するため，統計的流暢性及び文法的正確性の 2 つの観点で評価した．評価は，提案する 3 つのアプローチ（ビット分割，未完結化，文末生成制御）のそれぞれを個別に適用した場合と，全てのアプローチを適用した場合とで行った．さらに，LLM モデルの差が与える影響を評価し，提案手法の有効性を検証した．

5.1 評価環境

評価は，ローカル PC 上で英文のステゴ文章を生成することで行った．提案するステゴ文章生成を行う Python プログラムを実装し，英文のカバー文章・秘密文章を入力して英文のステゴ文章を生成した．LLM モデルには GPT-OSS-20B を用いた．計算リソースの効率的な利用および推論速度の向上を図るために同モデルを 4 ビットに量子化したモデルを採用した．量子化モデルの実行および推論を行うためのエンジンには llama.cpp[*1]を用いた．

評価では，1 個の秘密文章と 10 個のカバー文章を用意し，10 個のステゴ文章を生成して以下の 5 つの手法の流暢性を比較した．

(1) 既存手法：
文献 [5] で提案されている従来の LLM ステガノグラフィ手法である．3 節で示した手順でステゴ文章を生成した．

(2) アプローチ 1：
4.3 節で示したアプローチ 1 のみを適用する手法である．秘密文章から得られた確率順位系列 R^S を k ビットごとに分割して新たな確率順位系列 $R^{S'}$ を作成し，カバー文章から得られた確率順位系列 R^C と連結してステゴ文章を生成した．本稿の評価では分割ビット数 $k = 4$ とした．

(3) アプローチ 2：
4.4 節で示したアプローチ 2 のみを適用する手法である．既存手法と同様にステゴ文章を生成するが，確率順位系列 R^* が終端まで到達した後も文末記号のトークンが選択されるまでステゴ文章の生成を継続した．R^* の終端到達後の生成では，アプローチ 1 の k を用いて 2^k 未満の確率順位のトークンをランダムに選択した．

(4) アプローチ 3：
4.5 節で示したアプローチ 3 のみを適用する手法である．元のカバー文章の末尾に「接続詞 + コンマ」を連結した文章を新たなカバー文章とし，3 節で示した手順でステゴ文章を生成した．10 個のカバー文章にはそれぞれ異なる 10 種類の接続句（）を連結した．

(5) 提案手法：
提案手法する 3 つのアプローチのすべてを適用する手法である．カバー文章はアプローチ 3 と同様のものを用いた．

なお，評価に先立って生成したすべてのステゴ文章から秘

[*1] https://github.com/ggml-org/llama.cpp

表 1　生成されたステゴ文章の例．ステゴ文章の太字は秘密文章から生成された部分を示している．提案手法のカバー文章の太字はアプローチ 3 による既存手法との差を示している．

項目		内容
秘密文章		The official announcement regarding the merger is scheduled for next Friday. In preparation for that announcement, you may want to consider starting to sell your shares now.
既存手法 [5]	カバー文章	Here are some ideas for spending quality time with your family this summer.
	ステゴ文章	Here are some ideas for spending quality time with your family this summer. **This shouldn't turn them into little adults, but it could be done so they don't feel like it's their parents' time after school 'the kids didn't have**
提案手法（分割ビット数 $k = 4$）	カバー文章	Here are some ideas for spending quality time with your family this summer. **First,**
	ステゴ文章	Here are some ideas for spending quality time with your family this summer. First, **consider making use of a beautiful, spacious, and sunny location. If you have a backyard, you can have a picnic. Bring a nice blanket and a small table and set up a little picnic spot. If you have a pool and it's a great year, you could have a picnic around the pool or just set yourself at home with the best.**

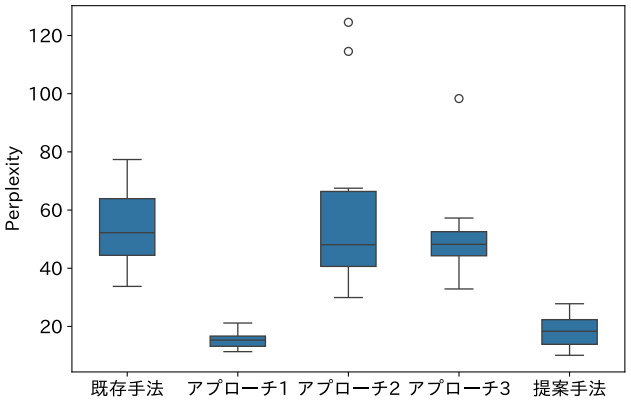

図 5　5 つの手法の PPL の分布．ボックス内の線は中央値を，ボックスの上下は四分位範囲を，エラーバーは残りの分布を，白丸は外れ値をそれぞれ示している．

密文章を正しく復元できることを確認した．

5.2　統計的流暢性

統計的流暢性は，LLM が文章をどの程度自然に予測できるかを測定する指標である Perplexity（PPL）によって評価した．PPL は言語モデルがテキストをどの程度自然に予測できるかを示す指標であり，1 以上の値を取り，値が小さいほどモデルにとって予測しやすい，すなわち流暢な文章であることを意味する [6]．

トークン長 Z の文章 $W = (w_1, w_2, \ldots, w_Z)$ に対する PPL は以下の式で定義される．

$$\mathrm{PPL}(W) = \exp\left(-\frac{1}{Z}\sum_{i=1}^{Z}\log P\left(w_i \mid P_{\mathcal{M}}(w_{<i})\right)\right) \quad (14)$$

ここで，$P\left(w_i|P_{\mathcal{M}}(w_{<i})\right)$ は確率分布 $P_{\mathcal{M}}(w_{<i})$ においてトークン w_i が選択される条件付き確率を表している．

図 5 に，5 つの手法の PPL の分布を示す．既存手法，アプローチ 1,2,3，提案手法の PPL の中央値は，それぞれ 52.2，15.3，48.1，48.2，18.4 である．図より，以下のことが分かる．

(1) アプローチ 1 と提案手法では，すべての文章で既存手法よりも PPL が低下した．既存手法と比べ，PPL はアプローチ 1 で約 1/3.4 倍，提案手法で約 1/2.8 倍となった．提案するビット分割による流暢性向上手法によって，既存手法よりも統計的流暢性を向上できたと言える．

(2) 提案手法はアプローチ 1 よりも PPL が若干大きい傾向にある．式 (14) より，いかなるトークンが選ばれようとも $P\left(w_i \mid P_{\mathcal{M}}(w_{<i})\right) < 1$ であるため，Z が大きいほど，すなわち文章が長いほど PPL は大きくなる．アプローチ 1 と比べて提案手法のステゴ文章はアプローチ 2 の効果により長くなるため，PPL が増加したと考えられる．

(3) アプローチ 2 及び 3 は，既存手法と比べて大きな PPL の変化は確認できなかった．アプローチ 2 及び 3 は単独で用いても PPL を大きく低下させる効果は無かったと言える．

以上の結果から，提案手法，特にアプローチ 1 によって既存手法と比べて統計的流暢性を向上できたことが確認された．

5.3　文法的正確性

文法的正確性は，オープンソースの校正エンジンである LanguageTool を用い，以下で定義される *GrammarPenaltyScore*(*GPS*) [7] 及びエラー率（100 トークンあたりの平均エラー数）によって評価した．*GPS* は 1 に近いほど文法的誤りが少ないことを示している．

$$GPS = 1 - \frac{N_{\mathrm{mistakes}}}{N_{\mathrm{words}} + 1} \quad (15)$$

ここで，N_{mistakes} は LanguageTool によって検出されたエラーの総数，N_{words} はステゴ文章の総単語数である．エラーには，文法エラー以外にもスペル，冠詞，時制，主語動詞一致，前置詞，句読点，大小文字，スタイル系の指摘も含まれる．分母の $+1$ はエラー数に対する感度を調整す

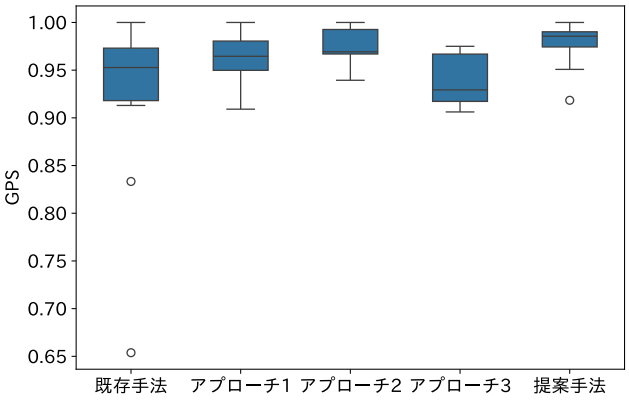

図 6 5 つの手法の GPS の分布

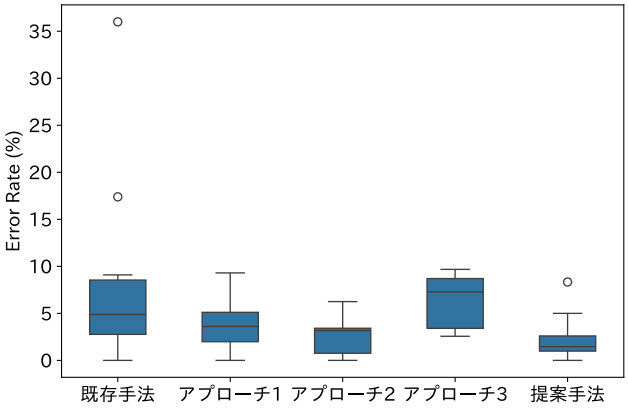

図 7 5 つの手法のエラー率の分布

るためのものである．*GPS* は文章の長さに依存するため，文章の長さに依存しない評価指標としてエラー率についても評価した．

図 6 及び**図 7** に，5 つの手法の GPS 及びエラー率の分布をそれぞれ示す．既存手法，アプローチ 1,2,3，提案手法の GPS の中央値は，それぞれ 0.95，0.96，0.97，0.93，0.99，エラー率の中央値は，それぞれ 4.9%，3.6%，3.2%，7.3%，1.5% である．図 6，図 7 より，以下のことが分かる．

(1) 5 つの手法の中で，提案手法の GPS がもっとも大きく，提案手法のエラー率がもっとも小さかった．提案手法は既存手法よりもエラー率を約 1/3.3 倍に低下させた．3 つのアプローチを組み合わせることで，文法的な正確性が向上したと言える．

(2) 提案手法の方がアプローチ 1 よりも大きい GPS，小さいエラー率となった．アプローチ 1 とアプローチ 2,3 を組み合わせることで文法的正確性がさらに向上したと言える．

(3) アプローチ 3 は既存手法と比べて GPS・エラー率を劣化させた．文章先頭は概してトークン確率順位が大きくなる傾向があり，カバー文章を未完結化したことで文章途中で大きい確率順位のトークンが選択され，文法的な正確性が低下したと考えられる．

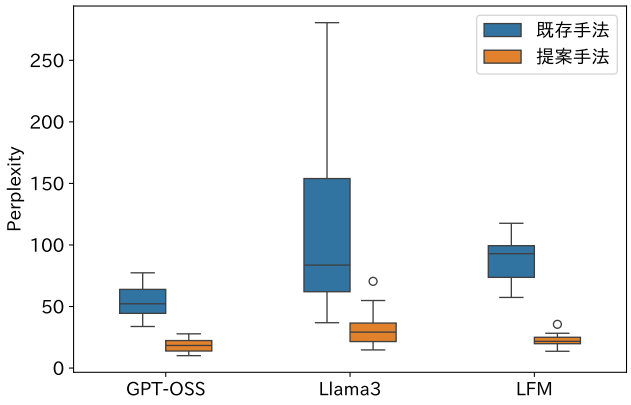

図 8 異なる LLM を用いた場合の既存手法 [5] と提案手法の PPL の分布

表 2 異なる LLM を用いた場合の既存手法 [5] と提案手法の PPL の中央値

GPT-OSS	既存手法	52.2
	提案手法	18.4
LFM	既存手法	92.9
	提案手法	21.7
Llama3	既存手法	83.7
	提案手法	29.2

以上より，3 つのアプローチを組み合わせた提案手法によって文法的正確性を向上できたことが確認された．

5.4 LLM モデルの影響

異なる LLM モデルを用いた場合の提案手法の有効性を検証するため，規模や構造の異なる以下の 3 つの LLM モデルを用いた場合で，既存手法と提案手法の PPL を比較した．

- GPT-OSS-20B
- Meta-Llama-3-8B
- LFM2.5-1.2B

図 8 に，異なる LLM を用いた場合の既存手法 [5] と提案手法の PPL の分布を示す．**表 2** に各 LLM，各手法の PPL の中央値を示す．図より，3 つの LLM モデルすべてにおいて，提案手法によって既存手法よりも PPL が低下したことが分かる．表 2 より，GPT-OSS，LFM，Llama3 では，それぞれ提案手法によって既存手法よりそれぞれ 1/2.8 倍，1/4.3 倍，1/2.9 倍となった．3 つの LLM モデルで確認した範囲では，提案手法は LLM モデルによらず統計的流暢性を向上させることを確認した．

6 おわりに

本稿では任意の秘密文章から流暢なステゴ文章を生成する LLM ステガノグラフィ手法を提案した．提案手法では，トークン確率順位のビット分割，文末までの生成制御やカバー文章の未完結化を行うことで，従来手法の課題であった生成文章の唐突な途切れや文脈の不自然さを解消した．

評価実験の結果，Perplexity（PPL）を用いた流暢性の評価および LanguageTool を用いた文法評価において，従来手法と比較して高い流暢性と文法的な正確性を維持できることを示した．提案した 3 つのアプローチを用いることで，10 個のステゴ文章の PPL を 1/2.8 倍に低下させることができた．校正エンジンによる文法評価において，エラー率を 1/3.3 倍に低下させることができた．

さらに，規模や構造の異なる 3 つの LLM を用いた検証により，本手法が特定のモデルに依存せず，高い汎用性を有することを確認した．

参考文献

[1] Subramanian, N., Elharrouss, O., Al-Maadeed, S. and Bouridane, A.: Image Steganography: A Review of the Recent Advances, *IEEE Access*, Vol. 9, pp. 23409–23423 (2021).

[2] Wu, Z., Guo, J., Zhang, C. and Li, C.: Steganography and Steganalysis in Voice over IP: A Review, *Sensors (Basel, Switzerland)*, Vol. 21, No. 4, p. 1032 (2021).

[3] J. Wu, Z. Wu, Y. Xue, J. Wen, and W. Peng: "Generative Text Steganography with Large Language Model," Proceedings of the 32nd ACM International Conference on Multimedia, pp. 10345–10353 (2024).

[4] 滝澤修, 松本勉, 中川裕志, 村瀬一郎, 牧野京子, "改行位置の調整によるドキュメントへの情報ハイディング," 情報通信研究機構季報, Vol. 51, Nos. 1/2, pp. 153–169 (2005).

[5] Norelli, A. and Bronstein, M.: LLMs can hide text in other text of the same length, *arXiv:2510.20075* (2025).

[6] A. Wuhrmann, A. Kucharavy, and A. Kucherenko: "Low-Perplexity LLM-Generated Sequences and Where To Find Them," Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pp. 774–783 (2025).

[7] Gavara Likhitha and B. Prajna, "AUTOMATED GRADING USING MACHINE LEARNING," *International Journal of Engineering Technology Research & Management (IJETRM)*, Vol. 09, Issue 07, pp.630–634 (2025).

[8] 持橋大地：大規模言語モデル（LLM）とロボティクス，日本ロボット学会誌，Vol. 40, No. 10, pp. 863–866 (2022).

[9] Jianxin Gao, Ruohan Lei, and Wanli Peng. "Text Steganography with Dynamic Codebook and Multimodal Large Language Model." *arXiv preprint arXiv:2604.20269*, (2026).

[10] R. Yan and Y. Murawaki, "Low-Overhead Disambiguation for Generative Linguistic Steganography via Tokenization Consistency," in *Proceedings of the 31st Annual Conference of the Association for Natural Language Processing*, pp. 2053–2058, March 2025.