УДК 004.056

СОКРЫТИЕ В СМЫСЛЕ: СЕМАНТИЧЕСКОЕ КОДИРОВАНИЕ ДЛЯ ГЕНЕРАТИВНО-ТЕКСТОВОЙ СТЕГАНОГРАФИИ

О. Ю. Рогов[1] [0000-0001-9672-2427], Д. Е. Инденбом[2] [0009-0001-9444-6075], Д. С. Корж[3] [0009-0000-6614-120X], Д. В. Пугачёва[4] [0000-0002-4285-1001], В. А. Воронов[5] [0000-0003-3835-6144], Е. В. Тутубалина[6] [0000-0001-7936-0284]

[1, 3, 4, 6]*Институт искусственного интеллекта, г. Москва, Россия*

[1, 2, 5]*Московский физико-технический институт, г. Долгопрудный, Россия*

[1, 3]*Московский технический университет связи и информатики, г. Москва, Россия*

[6]*Высшая школа экономики, г. Москва, Россия*

[6]*Казанский (Приволжский) федеральный университет, г. Казань, Россия*

[1]rogov@airi.net, [2]indenbom.de@phystech.edu, [3]korzh@airi.net, [4]daria.pugacheva@skoltech.ru, [5]v-vor@yandex.ru, [6]tutubalina@airi.net

Аннотация

В статье предложена новая система для генерации стеганографического текста, скрывающая двоичные сообщения в семантически связном естественном языке с помощью скрытого пространства, обусловливающего большие языковые модели (LLM). Секретные сообщения сначала кодируются в непрерывные векторы с помощью обученного отображения двоичного кода в скрытое пространство, которое используется для управления генерацией текста посредством донастройки префикса. В отличие от предыдущих методов стеганографии на уровне токенов или синтаксиса, наш метод позволяет избежать явной манипуляции словами и вместо этого работает полностью в скрытом семантическом пространстве, что обеспечивает более плавные и менее заметные результаты. На стороне получателя скрытое представление восстанавливается из сгенерированного текста и декодируется обратно в исходное сообщение.
В качестве ключевого теоретического вклада мы предоставляем гарантию надежности: если восстановленный скрытый вектор находится в пределах ограниченного расстояния от изначального, обеспечивается точное восстановление сооб-

щения, причем граница определяется константой Липшица декодера и минимальным отступом логитов. Этот формальный результат предлагает принципиальный подход к компромиссу между надежностью и емкостью в скрытых стеганографических системах. Эмпирическая оценка как на синтетических данных, так и в практических предметных областях, таких как отзывы на Amazon, показывает, что наш метод достигает высокой точности восстановления сообщений (выше 91%), высокую плавность текста и конкурентоспособную емкость до 6 бит на элемент предложения, сохраняя при этом устойчивость к нейронному стегоанализу. Эти результаты демонстрируют, что генерация со скрытым условием предлагает безопасный и практичный путь для встраивания информации в современные LLM.

***Ключевые слова**: стеганография, семантическое кодирование, языковые модели, донастройка префиксов, граф знаний, генерация естественного языка, скрытое обусловливание, нейронный стегоанализ.*

ВВЕДЕНИЕ

Способность скрытно встраивать информацию в естественный язык играет ключевую роль в безопасной коммуникации и цифровых водяных знаках. Традиционные методы стеганографии работают на уровне символов, слов или синтаксиса, часто вводя статистические артефакты или заметные возмущения в сгенерированный текст. С широким распространением мощных больших языковых моделей (LLM) появилась новая возможность встраивать информацию в сам смысл языка.

Мы представляем новую структуру для семантической стеганографии, которая скрывает сообщения не в выборе токенов на поверхностном уровне, а в скрытой семантической структуре.

Наш метод кодирует двоичные сообщения в виде плотных путей, каждый из которых представляет собой связную концептуальную структуру (например, «астронавт», «исследует», «планета», «удивлен»). Эти пути отображаются в непрерывные скрытые векторы с помощью кодировщика, а затем проецируются в пространство входных векторов LLM. Посредством донастройки префикса (prefix

tuning) полученный вектор обусловливает языковую модель генерировать плавные, похожие на человеческие предложения, которые сохраняют заданную семантику.

На этапе декодирования сгенерированный текст анализируется с помощью распознавания именованных сущностей (Named Entity Recognition — NER) и семантической маркировки ролей для восстановления исходного концептуального пути. Затем этот восстановленный путь отображается обратно в скрытое пространство и декодируется в исходное сообщение с помощью обратного отображения графа. Поскольку наш метод использует внутреннее семантическое выравнивание LLM, он позволяет избежать прямой манипуляции токенами и остается устойчивым в условиях «черного ящика», когда внутренние логиты или распределения выборки недоступны.

Мы проверили наш подход на синтетическом и открытом наборе структурированных текстов, продемонстрировав конкурентоспособную битовую емкость (5–6 бит на семантическую единицу), высокую лингвистическую естественность и устойчивость к современным системам стегоанализа. Насколько нам известно, это первая стеганографическая структура, которая согласовывает кодирование графа знаний с префиксным обусловливанием LLM для плавной и безопасной генерации текста.

Помимо эмпирических результатов, мы предоставляем формальные гарантии надежности [1] для восстановления сообщений в скрытом пространстве. В частности, мы анализируем условия, при которых двоичные сообщения, встроенные в непрерывные скрытые векторы, могут быть надежно декодированы после возмущений, например, возникающих во время генерации или извлечения. Наш анализ дает жесткие ограничения на допустимую величину возмущения с точки зрения константы Липшица декодера и отступа логитов от порога принятия решения. Этот результат устанавливает принципиальный компромисс между стабильностью скрытого кодирования и точностью декодирования и может послужить основой для будущих разработок доказательно безопасных или сертифицировано устойчивых стеганографических систем.

1.ЛИТЕРАТУРНЫЙ ОБЗОР

1.1. Генеративно-текстовая стеганография

Генеративно-текстовая стеганография использует предварительно обученные языковые модели для создания естественно выглядящих текстов, которые скрывают секретную информацию с помощью механизма стеганографического отображения. В таких системах качество как языковой модели, так и стратегии отображения играет критическую роль в обеспечении уровня скрытности и возможности восстановления данных.

В 2012 г. Эрнан Моральдо [2] предпринял одну из первых попыток создания генератора текста с встроенным зашифрованным сообщением. Для этого было предложено использовать языковую модель на основе цепи Маркова. В описанном методе в рамках марковской цепи анализируется корпус текстов для определения вероятностей переходов между токенами (словами) в пределах одного предложения. На основе этих вероятностей выбор каждого следующего генерируемого токена автоматически соотносится с кодируемой группой токенов. И каждый последующий переход сужает выбор кодируемых единиц информации вплоть до одной конкретной. Однако данный метод упирается в ограниченную емкость знаний марковской цепи, что приводит к неестественности генерируемого текста.

Позднее в работе Фанга и др. [3] был представлен новый основополагающий подход, в котором словарь *V* разделяется на 2^b непересекающихся групп $[V_1, V_2, \dots, V_{2^b}]$ для кодирования *b*-битных сегментов секретного сообщения. Во время генерации выбирается токен с наибольшей вероятностью в соответствующей группе. Последующие исследования Янга и др. [4, 5] продемонстрировали, что более совершенные языковые модели, такие как LSTM и BERT, значительно повышают как естественность, так и безопасность генерации стеганографического текста.

Зиглер и др. [6] и Дай и др. [7] применили GPT-2 в качестве базовой языковой модели и ввели отдельные стеганографические отображения, адаптированные к ее распределению токенов. Их результаты показали, что выбор функции

отображения оказывает значительное влияние на заметность и емкость стеганографических систем.

С точки зрения криптографии, Чжан и др. [8] предложили адаптивную динамическую группировку (Adaptive Dynamic Grouping — ADG), доказательно безопасный подход, который рекурсивно встраивает секретные биты посредством адаптивной группировки токенов в словаре. Из последних исследований отметим работу Динга и др. [9], где авторы представили метод Discop, который сохраняет исходное распределение токенов, копируя его в процессе встраивания, что обеспечивает высокие показатели незаметности и безопасности.

1.2. Большие языковые модели

Большие языковые модели (Large Language Model — LLM) продемонстрировали высокую эффективность в широком спектре генеративных задач. В недавних работах было исследовано их использование для генерации различных типов структурированных или размеченных данных, включая табличные записи [10], тройки для графов знаний [11] и пары предложений [12].

Большинство ранних подходов полагалось на простые классовые префиксы и zero-shot промпты для генерации данных в определенной области. Для улучшения устойчивости SuperGen [13] и ZeroGen [14] использовали LLM для генерации обучающих данных для задач классификации текста и включили устойчивые к шуму методы обучения [15] для устранения несоответствий в сгенерированных метках. SunGen [16] еще больше улучшил генерацию за счет определения весов, соответствующих качеству данных, для взвешенного использования синтетических примеров во время обучения, чтобы повысить общую эффективность.

Параллельно с этим недавние достижения в области промпт-инжиниринга сделали генерацию на основе LLM более контролируемой. Чэнь и др. [17] предложили подход мягкой донастройки промптов, применимый к LLM типа «белый ящик» с доступом к ключу случайной генерации. Ю и др. [18] расширили эту идею на окружения типа «черный ящик» и API к LLM (например, ChatGPT), продемонстрировав, что высококачественная генерация данных может быть достигнута без размеченных примеров или доступа к внутренним компонентам модели.

2. МЕТОД

2.1. Вводная информация

На рис. 1 представлена общая схема работы предлагаемого метода стеганографии. Ниже последовательно описывается каждый элемент системы.

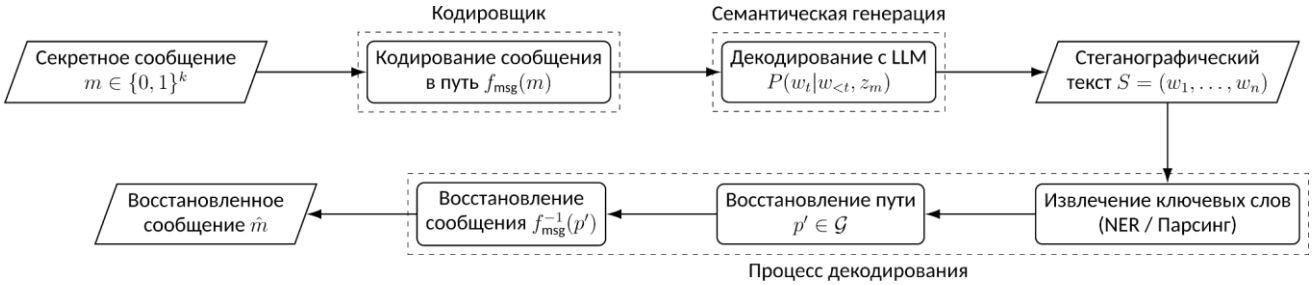

Рис. 1. Схема предлагаемого метода

Пусть $m \in \{0,1\}^k$ — секретное двоичное сообщение фиксированной длины k. Сообщение отображается в скрытый вектор $z_m \in R^d$ с помощью детерминированного или обучаемого кодировщика $f_{\text{enc}}: \{0,1\}^k \to R^d$. Это кодирование выполняет роль условного сигнала для генерации и сохраняется согласованным между отправителем и получателем. Для эффективной работы с языковой моделью (Language Model — LM) вектор z_m может быть спроецирован в пространство эмбеддингов декодера с помощью небольшого многослойного перцептрона (Multi-Layer Perceptron — MLP).

Скрытое обусловливание. Теперь рассмотрим LM, которая генерирует последовательность $S = (w_1, \ldots, w_n)$ из словаря D с помощью моделирования условных вероятностей:

$$P(S \mid z_m) = \prod_{t=1}^{n} P(w_t \mid w_{<t}, z_m),$$

где z_m — скрытый вектор, полученный из двоичного сообщения m. Данное обусловливание реализуется посредством донастройки префикса или прямого добавления эмбеддинга в скрытое состояние модели.

Генерация стеганографического текста. Во время генерации скрытый вектор z_m используется для направления модели на создание плавного текста, в ко-

торый неявно встроено сообщение. Модель обучается таким образом, чтобы изменения в z_m влияли на генерацию восстанавливаемым и декодируемым образом, без явного добавления битов сообщения в токены. Результатом является естественно выглядящее предложение *S*, которое кодирует *m* в своей скрытой семантической структуре.

Декодирование. Получив сгенерированное стеганографическое предложение S, получатель использует ту же языковую модель, чтобы получить соответствующий скрытый вектор $\hat{z}$. Затем применяется функция декодера $f_{dec}: R^d \to \{0,1\}^k$ для восстановления сообщения:

$$\widehat{m} = f_{\mathrm{dec}}(\hat{z}),$$

где $\widehat{m}$ — восстановленная битовая строка. Для правильного декодирования восстановленный скрытый вектор должен оставаться в пределах ϵ-шара исходного закодированного вектора $\|z_m - \hat{z}\|_2^2 \leq \epsilon$.

Система оптимизирована целиком (end-to-end) для достижения максимального качества текста при сохранении точного восстановления сообщения. В процессе обучения используется комбинированная функция потерь, включающая стандартное слагаемое, отвечающее за моделирование языка, и компонент, проверяющий восстановление данных:

$$L = L_{\mathrm{NLL}} + \lambda \cdot \|f_{\mathrm{dec}}(\hat{z}) - m\|_1, \qquad (1)$$

где L_{NLL} — отрицательный логарифм правдоподобия сгенерированного текста, а λ — коэффициент регуляризации, балансирующий плавность и восстанавливаемость.

2.2. Гарантии надежности

Лемма 1 (Скрытая надежность). *Пусть* $m \in \{0,1\}^k$ *— двоичное сообщение, закодированное в скрытый вектор* $z_m \in R^d$. *Предположим, что получатель принимает некоторое приближение* $\hat{z}$, *удовлетворяющее* $\| \hat{z} - z_m \|_2 \leq \delta$.

Пусть $f_{\mathrm{dec}}: R^d \to R^k$ *— декодер, являющийся* L*-липшицевым отображением*. Следовательно,

$$\|f_{\mathrm{dec}}(z_1) - f_{\mathrm{dec}}(z_2)\|_\infty \leq L\|z_1 - z_2\|_2 \text{ для любого } z_1,\ z_2 \in R^d.$$

Предположим, что f_{dec} генерирует логиты, такие что для истинного скрытого z_m:

$$\mathrm{round}\big(f_{\mathrm{dec}}(z_m)\big) = m, \quad \min_{1\le i\le k} |[f_{\mathrm{dec}}(z_m)]_i - 0.5| \ge \eta > 0,$$

где η — минимальный битовый отступ. Если $\delta < \eta/L$, то декодированное сообщение в точности соответствует m:

$$\mathrm{round}\big(f_{\mathrm{dec}}(\hat{z})\big) = m.$$

Доказательство. По условию Липшица и условию восстановления,

$$\max_{1\le i\le k} |[f_{\mathrm{dec}}(\hat{z})]_i - [f_{\mathrm{dec}}(z_m)]_i| = \| f_{\mathrm{dec}}(\hat{z}) - f_{\mathrm{dec}}(z_m) \|_\infty \le L\delta < \eta.$$

Это означает, что для каждого бита $i \in \{1, \dots, k\}$

$$|[f_{\mathrm{dec}}(\hat{z})]_i - [f_{\mathrm{dec}}(z_m)]_i| < \eta.$$

По предположению о битовом отступе в z_m:

- Если $m_i = 0$, то $[f_{\mathrm{dec}}(z_m)]_i \le 0.5 - \eta$, и, следовательно:

$$[f_{\mathrm{dec}}(\hat{z})]_i < [f_{\mathrm{dec}}(z_m)]_i + \eta \le (0.5 - \eta) + \eta = 0.5.$$

- Если $m_i = 1$, то $[f_{\mathrm{dec}}(z_m)]_i \ge 0.5 + \eta$, и, следовательно:

$$[f_{\mathrm{dec}}(\hat{z})]_i > [f_{\mathrm{dec}}(z_m)]_i - \eta \ge (0.5 + \eta) - \eta = 0.5.$$

В обоих случаях $[f_{\mathrm{dec}}(\hat{z})]_i$ находится строго на правильной стороне от 0.5. Следовательно,

$$\mathrm{round}([f_{\mathrm{dec}}(\hat{z})]_i) = m_i \text{ для любого } i$$

и $\mathrm{round}\big(f_{\mathrm{dec}}(\hat{z})\big) = m$.

3. ЭКСПЕРИМЕНТЫ И ИХ ОБСУЖДЕНИЕ

Мы оцениваем предложенную стеганографическую систему с помощью доступной легкой языковой модели, обусловленной на вектора двоичных сообщений. Наша цель — оценить способность системы генерировать плавный, незаметный стеганографический текст, при этом обеспечивая точное восстановление сообщения.

Каждое секретное сообщение m представляет собой строку из 64 бит, выбранную случайным образом из равномерного распределения, т. е.

$m \in \{0,1\}^{64}$. Двоичное сообщение отображается в скрытый вектор $z_m \in R^{128}$ с помощью обучаемой нейросети кодировщика f_{enc}, состоящей из двух полносвязанных слоев со скрытой размерностью 256 и функцией активации ReLU. Формально,

$$z_m = \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 m + b_1) + b_2).$$

Полученный скрытый вектор преобразуют и проецируют в последовательность эмбеддингов для префикса в форме (prefix_length, hidden_dim), где prefix_length = 10 и hidden_dim = 768.

В качестве базовой языковой модели мы используем GPT-Neo-125M от EleutherAI. Условный префикс реализуется с помощью адаптеров LoRA через библиотеку peft. Эмбеддинги префикса добавляются в начало потока эмбеддингов на вход модели, чтобы сформировать обусловленный контекст для генерации.

Обучение проводится с помощью оптимизатора AdamW со скоростью обучения (learning rate) 10^{-4}, размером батча, равным 16, и максимальной длиной последовательности, равной 64. Мы дообучаем модель в течение 5 эпох на синтетическом наборе данных из $5 \cdot 10^4$ пар сообщений и текстов. Модель обучается целиком (end-to-end), чтобы минимизировать функцию потерь из уравнения (1). Мы эмпирически подобрали $\lambda = 1.0$ для всех экспериментов.

Мы измеряем качество текста с помощью перплексии на модели GPT-2 и метрики BERTScore (F1) по отношению к эталонным текстам из тренировочного домена. Точность передачи (recovery) оценивается как процент двоичных сообщений, восстановленных со 100% битовой точностью с помощью f_{dec} из сгенерированного текста. Кроме того, битрейт рассчитывается как общее количество успешно декодированных битов, разделенных на количество сгенерированных токенов (bits per token).

Табл. 1 эмпирически подтверждает лемму 1, демонстрируя, что наш метод обеспечивает наджное восстановление сообщений в условиях практических ограничений. Практически идеальные показатели восстановления (91.2% для синтетических данных и 89.5% для данных Amazon) подтверждают, что декодер сохраняет достаточный отступ η в пространстве логитов, демонстрируя при этом выпол-

нение условия Липшица для константы L — в точности теоретического требования для соблюдения $\delta < \eta/L$. Более высокая восстанавливаемость синтетических данных соответствует их более низкой перплексии (17.3 *против* 26.7), что указывает на более предсказуемую генерацию текста, которая, по сути, сохраняет более крупные отступы η от порога принятия решений. Важно, что конкурентоспособные битрейты (3.9–3.4 битов на токен) доказывают, что эти гарантии надежности не мешают практической емкости. Оптимизируя геометрию скрытого пространства для максимизации η при минимизации L, мы достигаем условия стабильности леммы без потери плотности информации. Незначительное снижение точности восстановления в отзывах на Amazon отражает дополнительную сложность для обусловливания с помощью промпта, которая ужесточает соотношение η/L, но все же сохраняет запас прочности значительно выше порога ошибки.

Таблица 1. Количественное сравнение с базовыми решениями на синтетическом наборе данных и наборе отзывов на Amazon для фиксированного отображения словаря (Fixed Vocabulary Mapping – FVM), генератора на основе марковской цепи (Markov Chain Generator – MCG) и предлагаемого метода.

Method	**Perplexity ↓**	**F1 ↑**	**Recovery ↑**	**Bits per token ↑**
FVM–Synth	42.1	0.741	72.3%	2.0
FVM–Amazon	48.9	0.702	68.0%	1.8
MCG–Synth	35.8	0.780	76.1%	1.7
MCG–Amazon	43.2	0.735	71.4%	1.5
Ours–Synth	**17.3**	**0.890**	**91.2%**	**3.9**
Ours–Amazon	**26.7**	**0.870**	**89.5%**	**3.4**

Наш метод значительно превосходит оба базовых подхода: фиксированное отображение словаря (Fixed Vocabulary Mapping — FVM) и генератор на основе марковской цепи (Markov Chain Generator — MCG) — по всем оцениваемым показателям. По сравнению с FVM, который опирается на статичную группировку токенов без контекстной адаптации, наш подход показывает более чем на 20 баллов выше по метрике BERTScore F1 и улучшает коэффициент восстановления сообщений почти на 20% как на синтетических, так и на практических наборах данных. Аналогично, хотя MCG производит более естественные результаты, чем FVM,

благодаря своему вероятностному моделированию, ему не хватает семантической согласованности, и он предлагает только ограниченную битовую емкость (от 1.5 до 1.7 битов на токен), что намного ниже 3.4–3.9 битов на токен, достигаемых нашей моделью, направляемой в скрытом пространстве. Эти улучшения обеспечиваются генерацией, обусловленной в скрытном непрерывном пространстве, что позволяет достигнуть более богатой выразительности и более надежного встраивания сообщений без использования жестких лексических ограничений.

Лемма 1 устанавливает формальную гарантию надежности восстановления дискретных сообщений, которая согласуется с несколькими теоретическими подходами и расширяет их. Наше требование отступа η отражает принцип ограничивающих рамок в нейронном кодировании, где низкоразмерные представления достигают устойчивости, ограничивая динамику безопасными областями. Здесь η определяет именно такую область в пространстве логитов, гарантируя, что возмущения остаются в границах принятия решений.

Условие Липшица для константы L формализует стабильность при вмешательствах, что является краеугольным камнем современных унифицированных теорий надежности [8]. Рассматривая восстановление как относительную стабильность по отношению к l_2-ограниченным возмущениям, мы определяем надежность как целевую устойчивость при вмешательствах. Зависящий от данных параметр η позволяет избежать искусственного масштабирования 2^k, что отвечает критике чрезмерно консервативных теоретических ограничений. Эмпирические показатели восстановления из табл. 1 демонстрируют эксплуатационную жизнеспособность, показывая, что оптимизированные скрытые пространства обеспечивают компромисс между эффективностью и надежностью, предсказанный геометрической теорией обучения (geometric learning theory).

3.1. Ограничения

Хотя наша система демонстрирует высокую емкость, плавность и устойчивость при генерации стеганографического текста, она имеет ряд ограничений, которые открывают возможности для будущих исследований.

Во-первых, наш текущий декодер предполагает доступ к одной и той же языковой модели или модели с сопоставимой скрытой структурой. Это может

ограничить восстановление сообщений между вариантами моделей или системами на основе API, где внутренние представления различаются. Во-вторых, кодировщик и декодер между бинарным и скрытым пространствами обучаются на синтетических данных или на данных из конкретной области, которые могут плохо обобщаться для генерации в открытом домене без доменной адаптации. Наконец, наша теоретическая гарантия надежности предполагает фиксированный декодер с известной константой Липшица и разделяющим отступом. На практике надежная оценка этих параметров может быть нетривиальной, особенно для высокоразмерных скрытых пространств или в условиях интенсивных враждебных возмущений, что является предметом текущей работы.

Кроме того, хотя наш метод позволяет избежать возмущений на уровне токенов, он все же может создавать небольшие сдвиги в распределении, обнаруживаемые с помощью продвинутых моделей стегоанализа, обученных на скрытых или стилистических признаках. Наконец, наш подход требует умеренных вычислительных ресурсов для донастройки префикса и обусловливания на вектор, а расширение этого метода на очень большие модели (например, GPT5) или мультимодальные окружения может создать проблемы с масштабированием. Мы считаем, что эти ограничения могут быть устранены с помощью таких техник, как независимое от модели кодирование, динамическая калибровка декодера и состязательное дообучение для стеганографической инвариантности.

ЗАКЛЮЧЕНИЕ

Несмотря на добавленные ограничения на естественность и используемый словарный запас, модель сохраняет высокий показатель восстановления сообщений (89.5%) и приемлемый уровень перплексии. Битрейт несколько снижается из-за более длинных и связных предложений, необходимых для соответствия предметной области, но остается конкурентоспособным по сравнению с предыдущими разработками. Результаты подчеркивают масштабируемость и надежность нашего подхода к встраиванию на основе векторов в практических сценариях генерации текста.

Благодарности

Работа выполнена при частичной поддержке Программы стратегического академического лидерства Казанского (Приволжского) федерального университета («ПРИОРИТЕТ-2030»).

СПИСОК ЛИТЕРАТУРЫ

1. *Karimov E., Varlamov A., Ivanov D., Korzh D., and Rogov O.Y. Novel.* LossEnhanced Universal Adversarial Patches for Sustainable Speaker Privacy. — 2025. — 2505.19951.

2. *Moraldo H.H.* An Approach for Text Steganography Based on Markov Chains // ArXiv. 2014. Vol. abs/1409.0915.

3. *Fang T.*, *Jaggi M., Argyraki K.* Generating steganographic text with LSTMs // arXiv preprint arXiv:1705.10742. 2017.

4. *Yang Z.-L., Guo X.-Q., Chen Z.-M., Huang Y.-F., Zhang Y.-J*. RNN-stega: Linguistic steganography based on recurrent neural networks // IEEE Transactions on Information Forensics and Security. 2018. Vol. 14, No. 5. P. 1280–1295.

5. *Yang Z.-L., Zhang S.-Y., Hu Y.-T., Hu Z.-W., Huang Y.-F.* VAE-Stega: linguistic steganography based on variational auto-encoder // IEEE Transactions on Information Forensics and Security. 2020. Vol. 16. P. 880–895.

6. *Ziegler Z., Deng Y., Rush A. M*. Neural Linguistic Steganography // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. P. 1210–1215.

7. *Dai F.Z., Cai Z.* Towards near-imperceptible steganographic text // arXiv preprint arXiv:1907.06679. 2019.

8. *Zhang S., Yang Z., Yang J., Huang Y.* Provably Secure Generative Linguistic Steganography// Findings of the AssociationforComputational Linguistics: ACLIJCNLP 2021. 2021. P. 3046–3055.

9. Ding J., Chen K., Wang Y., Zhao N., Zhang W., Yu N. Discop: Provably Secure Steganography in Practice Based on “Distribution Copies” // 2023 IEEE Symposium on Security and Privacy (SP) / IEEE Computer Society. 2023. P. 2238– 2255.

10. *Borisov V., Seßler K., Leemann T., Pawelczyk M., Kasneci G.* Languagemodels are realistic tabular data generators // arXiv preprint arXiv:2210.06280. 2022.

11. *Chia Y.K., Bing L., Poria S., Si L.* RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction // arXiv preprint arXiv:2203.09101. 2022.

12. *Schick T.*, *Schütze H.* Generating datasets with pretrained language models // arXiv preprint arXiv:2104.07540. 2021.

13. *Meng Y., Huang J., Zhang Y., Han J.* Generating training data with language models: Towards zero-shot language understanding // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 462–477.

14. *Ye J.*, *Gao J., Li Q., Xu H., Feng J., Wu Z., Yu T., Kong L.* Zerogen: Efficient zero-shot learning via dataset generation // arXiv preprint arXiv:2202.07922. 2022.

15. *Wang Y., Ma X., Chen Z., Luo Y., Yi J., Bailey J.* Symmetric cross entropy for robust learning with noisy labels // Proceedings of the IEEE/CVF international conference on computer vision. 2019. P. 322–330.

16. *Gao J., Pi R., Yong L., Xu H., Ye J., Wu Z., Zhang W., Liang X., Li Z., Kong L.* Self-guided noise-free data generation for efficient zero-shot learning // International Conference on Learning Representations (ICLR 2023). 2023.

17. *Chen D.*, *Lee C., Lu Y., Rosati D., Yu Z.* Mixture of Soft Prompts for Controllable Data Generation // arXiv preprint arXiv:2303.01580. 2023.

18. *Yu Y., Zhuang Y., Zhang J., Meng Y., Ratner A., Krishna R., Shen J., Zhang C.* Large language model as attributed training data generator: A tale of diversity and bias // arXiv preprint arXiv:2306.15895. 2023.

HIDING IN MEANING: SEMANTIC ENCODING FOR GENERATIVE TEXT STEGANOGRAPHY

O. Y. Rogov[1] [0000-0001-9672-2427], D. E. Indenbom[2] [0009-0001-9444-6075],
D. S. Korzh[3] [0009-0000-6614-120X], D. V. Pugacheva[4] [0000-0002-4285-1001],
V.A. Voronov[5] [0000-0003-3835-6144], E.V. Tutubalina[6] [0000-0001-7936-0284]

[1, 3, 4, 6]Artificial Intelligence Research Institute*, Moscow, Russia*
[1, 2 ,5]Moscow Institute of Physics and Technology, Dolgoprudny, Russia
[1, 3]Moscow Technical University of Communications and Informatics, Moscow, Russia
[6]HSE University, Moscow, Russia
[6]Kazan Federal University, Kazan, Russia

[1]rogov@airi.net, [2]indenbom.de@phystech.edu, [3]korzh@airi.net, [4]daria.pugacheva@skoltech.ru, [5]v-vor@yandex.ru, [6]tutubalina@airi.net

Abstract

We propose a novel framework for steganographic text generation that hides binary messages within semantically coherent natural language using latent-space conditioning of large language models (LLMs). Secret messages are first encoded into continuous vectors via a learned binary-to-latent mapping, which is used to guide text generation through prefix tuning. Unlike prior token-level or syntactic steganography, our method avoids explicit word manipulation and instead operates entirely within the latent semantic space, enabling more fluent and less detectable outputs. On the receiver side, the latent representation is recovered from the generated text and decoded back into the original message. As a key theoretical contribution, we provide a robustness guarantee: if the recovered latent vector lies within a bounded distance of the original, exact message reconstruction is ensured, with the bound determined by the decoder's Lipschitz continuity and the minimum logit margin. This formal result offers a principled view of the reliability–capacity trade-off in latent steganographic systems. Empirical evaluation on both synthetic data and real-world domains such as Amazon reviews shows that our method achieves high message recovery accuracy (above 91%), strong text fluency and competitive capacity up to 6 bits per sentence element while maintaining resilience against neural steganalysis. These findings demonstrate that latent

conditioned generation offers a secure and practical pathway for embedding information in modern LLMs.

Keywords*: steganography, semantic encoding, language models, prefix tuning, knowledge graphs, natural language generation, latent conditioning, neural steganalysis.*

REFERENCES

1. *Karimov E., Varlamov A., Ivanov D., Korzh D., and Rogov O.Y. Novel.* LossEnhanced Universal Adversarial Patches for Sustainable Speaker Privacy. — 2025. — 2505.19951.

2. *Moraldo H.H.* An Approach for Text Steganography Based on Markov Chains // ArXiv. 2014. Vol. abs/1409.0915.

3. *Fang T.*, *Jaggi M., Argyraki K.* Generating steganographic text with LSTMs // arXiv preprint arXiv:1705.10742. 2017.

4. *Yang Z.-L., Guo X.-Q., Chen Z.-M., Huang Y.-F., Zhang Y.-J*. RNN-stega: Linguistic steganography based on recurrent neural networks // IEEE Transactions on Information Forensics and Security. 2018. Vol. 14, No. 5. P. 1280–1295.

5. *Yang Z.-L., Zhang S.-Y., Hu Y.-T., Hu Z.-W., Huang Y.-F.* VAE-Stega: linguistic steganography based on variational auto-encoder // IEEE Transactions on Information Forensics and Security. 2020. Vol. 16. P. 880–895.

6. *Ziegler Z., Deng Y., Rush A. M*. Neural Linguistic Steganography // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. P. 1210–1215.

7. *Dai F.Z., Cai Z.* Towards near-imperceptible steganographic text // arXiv preprint arXiv:1907.06679. 2019.

8. *Zhang S., Yang Z., Yang J., Huang Y.* Provably Secure Generative Linguistic Steganography// Findings of the AssociationforComputational Linguistics: ACLIJCNLP 2021. 2021. P. 3046–3055.

9. Ding J., Chen K., Wang Y., Zhao N., Zhang W., Yu N. Discop: Provably Secure Steganography in Practice Based on “Distribution Copies” // 2023 IEEE Symposium on Security and Privacy (SP) / IEEE Computer Society. 2023. P. 2238– 2255.

10. *Borisov V., Seßler K., Leemann T., Pawelczyk M., Kasneci G.* Languagemodels are realistic tabular data generators // arXiv preprint arXiv:2210.06280. 2022.

11. *Chia Y.K., Bing L., Poria S., Si L.* RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction // arXiv preprint arXiv:2203.09101. 2022.

12. *Schick T.*, *Schütze H.* Generating datasets with pretrained language models // arXiv preprint arXiv:2104.07540. 2021.

13. *Meng Y., Huang J., Zhang Y., Han J.* Generating training data with language models: Towards zero-shot language understanding // Advances in Neural Information Processing Systems. 2022. Vol. 35. P. 462–477.

14. *Ye J.*, *Gao J., Li Q., Xu H., Feng J., Wu Z., Yu T., Kong L.* Zerogen: Efficient zero-shot learning via dataset generation // arXiv preprint arXiv:2202.07922. 2022.

15. *Wang Y., Ma X., Chen Z., Luo Y., Yi J., Bailey J.* Symmetric cross entropy for robust learning with noisy labels // Proceedings of the IEEE/CVF international conference on computer vision. 2019. P. 322–330.

16. *Gao J., Pi R., Yong L., Xu H., Ye J., Wu Z., Zhang W., Liang X., Li Z., Kong L.* Self-guided noise-free data generation for efficient zero-shot learning // International Conference on Learning Representations (ICLR 2023). 2023.

17. *Chen D.*, *Lee C., Lu Y., Rosati D., Yu Z.* Mixture of Soft Prompts for Controllable Data Generation // arXiv preprint arXiv:2303.01580. 2023.

18. *Yu Y., Zhuang Y., Zhang J., Meng Y., Ratner A., Krishna R., Shen J., Zhang C.* Large language model as attributed training data generator: A tale of diversity and bias // arXiv preprint arXiv:2306.15895. 2023.

СВЕДЕНИЯ ОБ АВТОРАХ

РОГОВ Олег Юрьевич — получил степень магистра в МГУ и степень кандидата наук в области математического моделирования и физики в Центре фотоники Российской академии наук. В настоящее время является соруководителем проекта машинного зрения в Университете Шарджи и руководителем исследовательской группы в AIRI. Его научные интересы включают эволюционные алгоритмы и глубокое обучение, слияние информации и принятие решений, а также основы проектирования и оценки производительности систем визуализации, обработку сигналов и обнаружение аномалий.

Oleg Yurievich ROGOV — received a Master's degree from MSU and a Ph.D. degree in mathematical modeling and physics at the Photonics Centre of the Russian Academy of Sciences. He is currently the co-PI of the medical vision project at the University of Sharjah and head of a research group at AIRI. His research interests include evolutionary algorithms and deep learning, information fusion and decision making, fundamentals of imaging system design and performance evaluation, signal processing, anomaly detection and estimation.

email: rogov@airi.net

ORCID: 0000-0001-9672-2427

ИНДЕНБОМ Дмитрий Евгеньевич — окончил бакалавриат и получил степень магистра по направлению «Прикладная математика и физика» в МФТИ. В настоящее время является аспирантом МФТИ и работает над диссертацией по теме «Методы защиты информации при использовании больших языковых моделей». Его научные интересы включают цифровую маркировку синтетических данных, скрытое шифрование сообщений в генерируемом тексте и интерпретацию работы языковых моделей.

Dmitrii Evgenievich INDENBOM — received a Bachelor's and a Master's degrees in applied mathematics and physics from MIPT. He is currently working on his dissertation on "Information security methods in the context of large language models" at the graduate school of MIPT. His research interests include digital watermarking of synthetic data, secret messages embedding in generated text, and interpretation of language models.

email: indenbom.de@phystech.edu

ORCID: 0009-0001-9444-6075

Корж Дмитрий Сергеевич — окончил бакалавриат в МФТИ и получил степень магистра в области наук о данных в Сколковском институте наук и технологий; продолжает научную работу в аспирантуре Сколтеха. В настоящее время является младшим научным сотрудником в группе «Доверенные и безопасные интеллектуальные системы» в AIRI и в «Лаборатории безопасного искусственного интеллекта» МТУСИ. Его научные интересы включают методы устойчивости нейронных сетей, безопасность алгоритмов глубокого обучения, прикладные задачи в звуковом домене.

Dmitrii Sergeevich Korzh — received a Bachelor's degree from MIPT and a Master's degree in data science from the Skolkovo Institute of Science and Technology, where he continues his scientific work in graduate school. He is currently a junior researcher in the Reliable and Secure Intelligent Systems group at AIRI and at the Laboratory of Secure Artificial Intelligence at MTUCI. His research interests include methods of robustness of neural networks, the security of deep learning algorithms, and applied problems in the audio domain.

email: korzh@airi.net
ORCID: 0009-0000-6614-120X

ПУГАЧЁВА Дарья Валерьевна — окончила бакалавриат в МФТИ, получила степени магистра по направлениям «Прикладные математика и физика» и «Математика и компьютерные науки» в МФТИ и в Сколковском институте наук и технологий, получила степень кандидата физико-математических наук, защитив диссертацию на тему «Лазерно-плазменное ускорение поляризованных заряженных частиц» в МФТИ. В настоящее время работает научным сотрудником в исследовательской группе прикладного NLP в AIRI. Ее научные интересы включают вопросы генерализации и устойчивости моделей для управления воплощенными агентами, комбинаторную оптимизацию, графовые нейронные сети.

Darya Valeryaevna PUGACHEVA — received a Bachelor's degree from MIPT and completed a Master's degrees in applied mathematics and physics and in mathematics and computer science at MIPT and at the Skolkovo Institute of Science and Technology. She obtained a PhD in physics and mathematics after defending her dissertation on "Laser–plasma

acceleration of polarized charged particles" at MIPT. She is currently a researcher in the Domain-specific NLP research group at AIRI. Her research interests include improving the generalization and robustness of models for the control of embodied agents, combinatorial optimization, and graph neural networks.

email: Daria.Pugacheva@skoltech.ru

ORCID: 0000-0002-4285-1001

ВОРОНОВ Всеволод Александрович — окончил Иркутский государственный университет, получил степень кандидата технических наук Институте проблем управления РАН. В настоящее время является заведующим лабораторией комбинаторной геометрии Кавказского математического центра Адыгейского государственного университета, работает старшим научным сотрудником лаборатории комбинаторных и геометрических структур Московского физико-технического института. Его научные интересы включают теорию графов, дискретную геометрию, динамические системы и машинное обучение.

Vsevolod Alexandrovich VORONOV — graduated from Irkutsk State University and received a PhD in technical sciences at the Institute of Control Sciences of RAS. He is the head of the Laboratory of Combinatorial Geometry at the Caucasus Mathematical Center of the Adyghe State University and a senior researcher in the Laboratory of Combinatorial and Geometric Structures at MIPT. His research interests include graph theory, discrete geometry, dynamical systems, and machine learning.

email: v-vor@yandex.ru

ORCID: 0000-0003-3835-6144

ТУТУБАЛИНА Елена Викторовна — получила степень кандидата физико-математических наук в Институте системного программирования им. В.П. Иванникова РАН и степень доктора компьютерных наук, защитив диссертацию на тему “Модели и методы автоматической обработки неструктурированных данных в биомедицинской области” в ВШЭ. В настоящее время является руководителем научной группы Domain-specific NLP в Институте AIRI, старшим научным сотрудником ИСП РАН и Казанского федерального университета. Ее научные интересы включают машинное обучение, обработку естественного языка, и исследование генерализации и устойчивости языковых моделей.

Elena Viktorovna TUTUBALINA — received a PhD in physics and mathematics at the Ivannikov Institute for System Programming of the Russian Academy of Sciences and the degree of Doctor of Computer Science after defending a dissertation on “Models and methods for automatic processing of unstructured data in the biomedical domain” at HSE. She currently leads the Domain-specific NLP research group at AIRI and is a Senior Researcher at the Ivannikov Institute for System Programming of the Russian Academy of Science and at the Kazan Federal University. Her research interests include machine learning, natural language processing, and studies of generalization and robustness of language models.

email: tutubalina@airi.net

ORCID: 0000-0001-7936-0284

Материал поступил в редакцию 14 октября 2025 года