



МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Филологический факультет

АКТУАЛЬНЫЕ ПРОБЛЕМЫ ЛИНГВИСТИКИ И ЛИТЕРАТУРОВЕДЕНИЯ

Сборник материалов X (XXIV)
Международной научно-практической
конференции молодых учёных
(13–15 апреля 2023 г.)

Выпуск 24

Томск
Издательство Томского государственного университета
2023

MINISTRY OF SCIENCE AND HIGHER EDUCATION
OF THE RUSSIAN FEDERATION
NATIONAL RESEARCH TOMSK STATE UNIVERSITY
Faculty of Philology

ACTUAL PROBLEMS OF LINGUISTICS AND LITERARY STUDIES

Proceedings of the X (XXIV)
International Scientific and Practical Conference
of Young Scientists
(April 13–15, 2023)

Issue 24

Tomsk
TSU Press
2023

УДК 81'1(082)

ББК 81

А43

Редакционная коллегия:

И.В. Тубалова (Томский государственный университет)
Т.А. Демешкина (Томский государственный университет)
З.И. Резанова (Томский государственный университет)
И.А. Айзикова (Томский государственный университет)
Н.Е. Никонова (Томский государственный университет)
В.С. Киселев (Томский государственный университет)
В.А. Суханов (Томский государственный университет)
С.В. Ермоленко (Томский государственный университет)
А.Г. Кожевникова (ответственный редактор;
Томский государственный университет)

А43 Актуальные проблемы лингвистики и литературоведения :
сборник материалов X (XXIV) Международной научно-
практической конференции молодых ученых (13–15 апреля
2023 г.) / отв. ред. А.Г. Кожевникова. – Томск : Издательство
Томского государственного университета, 2023. –
Вып. 24. – 664 с.

ISBN 978-5-907722-02-6

В сборнике представлены результаты научных исследований студентов, аспирантов, молодых ученых и преподавателей. Работы отражают многообразие актуальных проблем лингвистики, литературоведения, методики преподавания РКИ, теории перевода и издательского дела. Все статьи прошли процедуру обязательного рецензирования.

Для преподавателей вузов, учителей русского языка и литературы, студентов гуманитарных специальностей.

УДК 81'1(082)

ББК 81

ISBN 978-5-907722-02-6

© Томский государственный университет, 2023

UDC 81'1(082)
LBC 81
A43

Editorial advisory board:

I.V. Tubalova (National Research Tomsk State University)
T.A. Demeshkina (National Research Tomsk State University)
Z.I. Rezanova (National Research Tomsk State University)
I.A. Aizikova (National Research Tomsk State University)
N.E. Nikonova (National Research Tomsk State University)
V.S. Kiselev (National Research Tomsk State University)
V.A. Sukhanov (National Research Tomsk State University)
S.V. Ermolenko (National Research Tomsk State University)
A.G. Kozhevnikova (executive editor;
National Research Tomsk State University)

A43 Actual problems of linguistics and literary studies:
proceedings of the X (XXIV) International scientific and
practical conference of young scientists (April, 13-15 2023) / ed.
by A.G. Kozhevnikova. – Tomsk : TSU Press, 2023. –
Issue 24. – 664 p.

ISBN 978-5-907722-02-6

The book of proceedings presents the results of research work performed by students, postgraduate students, young scientists and teachers. The reports reflect the variety of advanced problems in linguistics, literary studies, teaching Russian as a foreign language, translation theory and publishing. All articles are peer-reviewed.

The book is of interest for university professors, school teachers of Russian language and literature, students of different humanitarian specialties.

UDC 81'1(082)
LBC 81

**ОСОБЕННОСТИ ПРИМЕНЕНИЯ МУЛЬТИЯЗЫЧНЫХ
ВРЕДОНОСНЫХ ЗАПРОСОВ НА ОСНОВЕ
СЛОВ-ГИБРИДОВ ДЛЯ ГЕНЕРАЦИИ ИЗОБРАЖЕНИЙ
НЕЙРОННОЙ СЕТЬЮ KANDINSKY 2.0¹**

Аишева Д.А.

Томский государственный университет, аспирант

**ADVERSARIAL ATTACKS ON KANDINSKY 2.0 IMAGE
GENERATION USING MULTI-LANGUAGE MACARONIC
PROMPTS**

Aisheva D.A.

Tomsk State University, postgraduate student

Вопросы безопасности нейронной сети Kandinsky 2.0 в контексте возможности применения к ней вредоносных запросов, сконструированных из слов на нескольких языках, не были ранее рассмотрены. Данная статья заполняет имеющийся пробел: представлены результаты репликации эксперимента по генерации изображений на основе мультязычных вредоносных запросов.

Ключевые слова: нейронные сети, Kandinsky 2.0, генерация изображений, вредоносные запросы, macaronic prompting

The security issues of the Kandinsky 2.0 neural network haven't yet been considered in the context of the possibility of adversarial attacks using macaronic prompts. In this article, we fill in the existing gap and present the results of the experiment on generating images based on macaronic prompts.

Keywords: neural networks, Kandinsky 2.0, image generation, adversarial prompts, macaronic prompting

Научный руководитель: З.И. Резанова, д-р филол. наук, зав. кафедрой общей, компьютерной и когнитивной лингвистики ФилФ ТГУ.

Активно исследуемый аспект изучения нейронных сетей – обеспечение безопасности их использования и нахождение уязвимостей [1. С. 2]. В частности, активно разрабатывается проблема подверженности текстовых нейронных сетей атакам при

¹ Статья подготовлена при поддержке Программы развития Томского государственного университета (Приоритет-2030).

помощи искажения вводимого текста. Вместе с тем данный вопрос практически не был исследован применительно к моделям нейронных сетей, генерирующих изображения на основе текста, несмотря на особенную актуальность, связанную с нарастающей скоростью производства таких моделей и слабыми механизмами защиты, основанными на поиске ключевых слов в запросе [2]. Не была ранее исследована эта тема и на материале нейронных сетей с большим количеством русскоязычных текстов в датасете.

Одной из значимых работ, в которой решается проблема уязвимости нейронных сетей, генерирующих изображения по вредоносным запросам на основе искажения текста, является статья R. Milliére «Adversarial Attacks on Image Generation with Made-Up Words», представляющая результат эксперимента по созданию и применению таких запросов к модели DALL-E.

Один из описанных в этой статье сценариев атак – *macaronic prompting* – основан на разбиении вредоносного текста, состоящего из слова, переведенного на несколько языков, представленных в модели, на токены (подстроки) и последующей конкатенации их в не имеющую смысла строку (слово-гибрид) [3]. Предполагается, что такой сценарий эксплуатирует сразу несколько особенностей модели – многоязычие, наличие подстрочного токенизатора и систему фильтрации запросов на основе ключевых слов.

Эти же свойства отличают модель Kandinsky от DALL-E и потому вызывают интерес к применению сценария *macaronic prompting* к Kandinsky 2.0. Помимо того, что модель Kandinsky отличается от DALL-E датасетом с большим количеством русскоязычных текстов, создатели модели подчеркивают разнообразие языков текстов в датасете [4], что открывает перспективы дальнейших исследований в области генерации запросов из иных, не упомянутых в оригинальной статье, языков. Ещё одно существенное отличие архитектуры Kandinsky – использование сразу двух токенизаторов. Один из них, XLMR, дает близкие представления для разных языков, другой токенизатор, mt5, отличается «хорошим пониманием сложных текстов» [4]. Особенностью также является слабая обеспеченность безопасности

моделей семейства Kandinsky; кроме того, в отличие от безопасности модели DALL-E, она совсем не описана.

В статье представлены результаты репликации эксперимента с применением к нейронной сети Kandinsky *masaronic prompting* и сравнение полученных результатов с результатами статьи, представившей метод.

В статье R. Milliére не описан точный алгоритм генерации *masaronic* запросов, однако упомянуты общие принципы их построения:

- выбранные слова переведены на несколько языков, представленных в корпусе модели;
- полученные слова разбиваются на токены токенизатором модели;
- токены сложены случайным образом в так называемый *masaronic* запрос.

Для создания запросов в нашем эксперименте был взят набор из 9 слов из статьи R. Milliére (*insect, lizard, butterfly, plane, firefighter, cliff, rabbit, education, exasperated*) и их переводы на 4 языка (немецкий, испанский, итальянский, французский).

Нами было создано по 4 запроса для каждого слова методом случайного сложения токенов, полученных при помощи 2 токенизаторов, после чего на основе каждого запроса сгенерировано по 4 изображения.

Далее запросы были проанализированы и разделены на группы по следующим критериям: соответствие образа в изображении запросу, устойчивость образа для запроса и объяснимость связи изображения и запроса. Связи между использованием разных алгоритмов токенизации для создания запроса и особенностями генерации по запросу не было найдено, поэтому результаты были объединены.

Группа 1 – изображения соответствуют оригинальному тексту, образ устойчиво появляется на всех изображениях по запросу, связь изображения и запроса объяснима.

Группа 2 – изображения не соответствуют оригинальному тексту, однако есть образ, устойчиво проявляющийся на всех изображениях по запросу, связь изображения и запроса объяснима.

Группа 3 – изображения не соответствуют оригинальному тексту, однако есть образ, устойчиво проявляющийся на всех изображениях по запросу, связь изображения и запроса при этом пока необъяснима.

Группа 4 – изображения не соответствуют оригинальному тексту, нет образа, устойчиво проявляющегося на всех изображениях по запросу, связь изображения и запроса необъяснима.

В группе запросов, по которым были сгенерированы изображения, соответствующие желаемому результату генерации запроса (группа 1), можно выделить следующие лингвистические особенности запросов: первые и последние токены в запросе соответствуют начальным и конечным токенам в переведенных словах (рис. 1).



Рис. 1. Изображение по запросу «feuerpompierleieremann»

Эта особенность подтверждается и при анализе следующей группы запросов – группы 2. Запросы из этой группы содержат в начале или конце подстроки, не соответствующие оригинальному

тексту, однако обладающие смыслом и значительно влияющие на генерацию изображения, что подтверждается устойчивостью образа внутри группы изображений. Так, например, по запросу «ferinsectkabiesin» (запрос на основе слова insect) были получены изображения папоротников, по запросу «tterillonfarschpapa» (запрос на основе слова butterfly) – характерные изображения пожилых мужчин (рис. 2).



Рис. 2. Изображение по запросам «ferinsectkabiesin» (слева)
и «tterillonfarschpapa» (справа)

Показателен пример запроса, для которого были сгенерированы устойчивые изображения прямоугольного объекта с изображением насекомого; начальная и конечная подстроки запроса – insect и kasetti (насекомое и кассета) (рис. 3).

Для большей части запросов из группы 3 требуется более глубокий анализ, направленный на изучение связи между изображениями и запросами. Так, для запроса «alingmetterfarposa» (запрос на основе слова butterfly) было получено устойчивое изображение шмеля на желтом цветке, однако лингвистический анализ не показал связи между запросом и изображением (рис. 4). Для запросов этой группы не удастся выделить характерные лингвистические особенности.



Рис. 3. Изображение по запросу «insectinchosferkasetti»

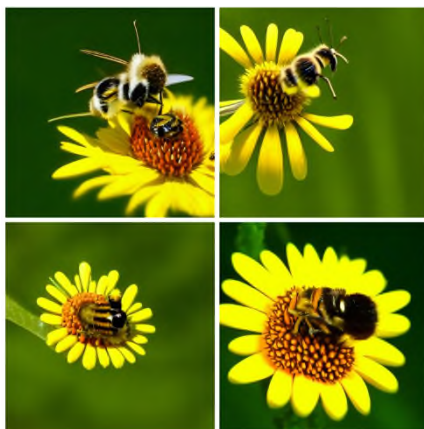


Рис. 4. Изображение по запросу «alingmetterfarposa»

Не представляется возможным выделить отличия запросов группы 3 от запросов группы 4, по которым на один запрос генерируются значительно отличающиеся и не наполненные смыслом изображения.

Полученные результаты более разнообразны, чем отображенные в статье, представившей метод, однако также приводят к выводу о

возможности успешного использования вредоносных запросов применительно к нейронным сетям для генерации изображений.

Представляется перспективным дальнейшее изучение данного метода на материале иных языков, представленных в модели, исследование иных методов создания вредоносных запросов путем искажения текста, а также генерация запросов на основе конкатенации токенов лексических единиц одного языка, лежащих внутри одного ЛСП.

Список источников

1. Chakraborty A. Adversarial Attacks and Defences: A Survey // arXiv. [Б. м.], 2018. URL: <https://arxiv.org/abs/1810.00069> (дата обращения: 01.02.2023).
2. DALL·E 2 Preview - Risks and Limitations // GitHub. [Б. м.], б. г. URL: <https://github.com/openai/dalle-2-preview/blob/main/system-card.md> (дата обращения: 01.02.2023).
3. Millièrè R. Adversarial Attacks on Image Generation with Made-Up Words // arXiv. [Б. м.], 2022. URL: <https://arxiv.org/abs/2208.04135> (дата обращения: 04.03.2023).
4. Kandinsky 2.0 – первая мультязычная диффузия для генерации изображений по тексту // Хабр. [Б. м.], 2006–2023. URL: <https://habr.com/ru/companies/sberbank/articles/701162/> (дата обращения: 22.03.2023).

doi: 10.17223/978-5-907572-02-7-2023-27

МОДЕЛИРОВАНИЕ ДИАЛОГОВЫХ ДЕРЕВЬЕВ В ОБРАЗОВАТЕЛЬНОЙ ВИДЕОИГРЕ «АРКА КРАЯ»

Селиванов Д.Д.

Томский государственный университет, студент

MODELLING DIALOGUE TREES IN THE EDUCATIONAL VIDEO GAME «ARKA KRAJA»

Selivanov D.D.

Tomsk State University, student

В статье рассматривается механика диалоговых деревьев в аспекте задач разработчика видеоигр: функции и принципы её построения, а также