

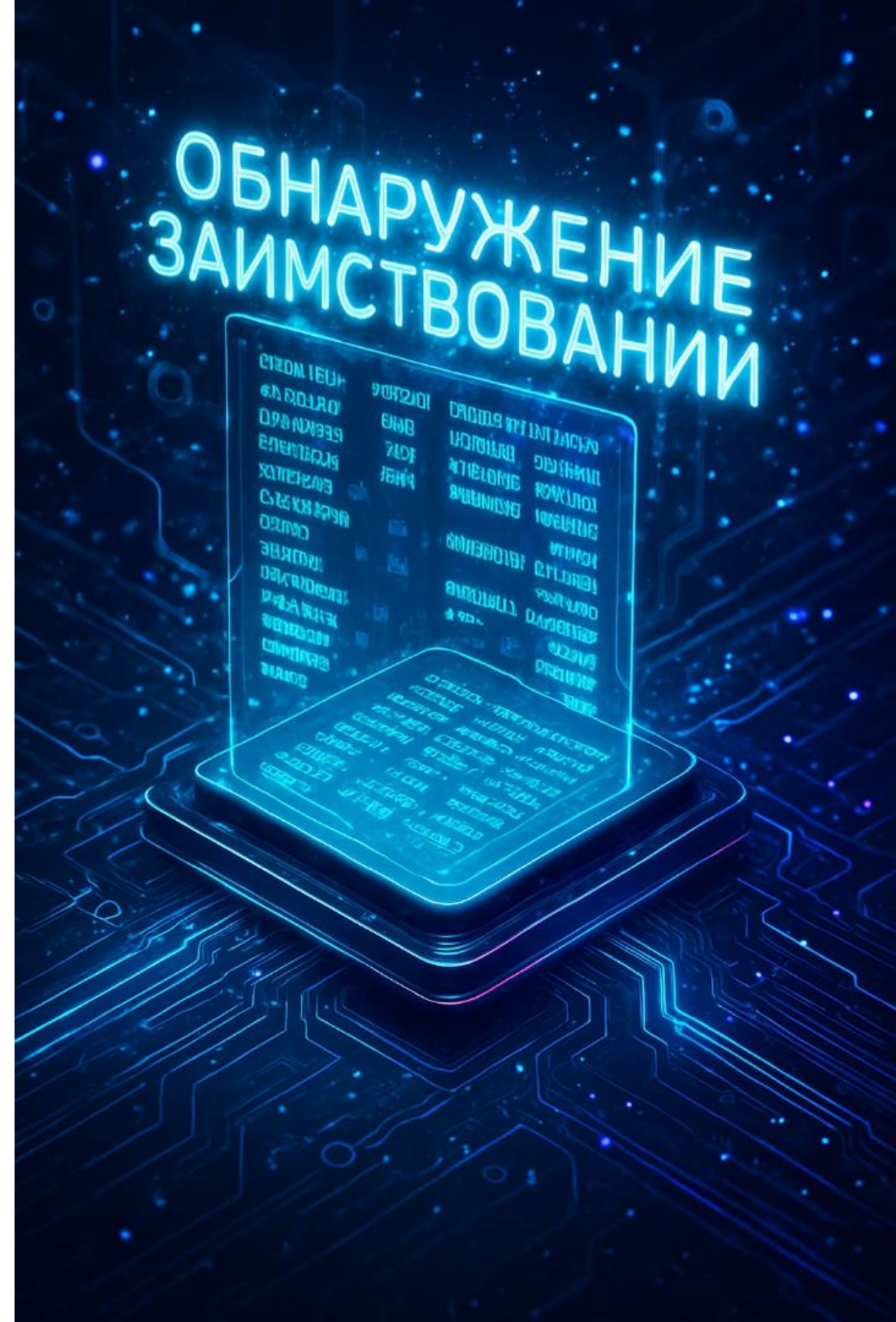
Задачи и алгоритмы систем обнаружения заимствований

Юрий Викторович Чехович

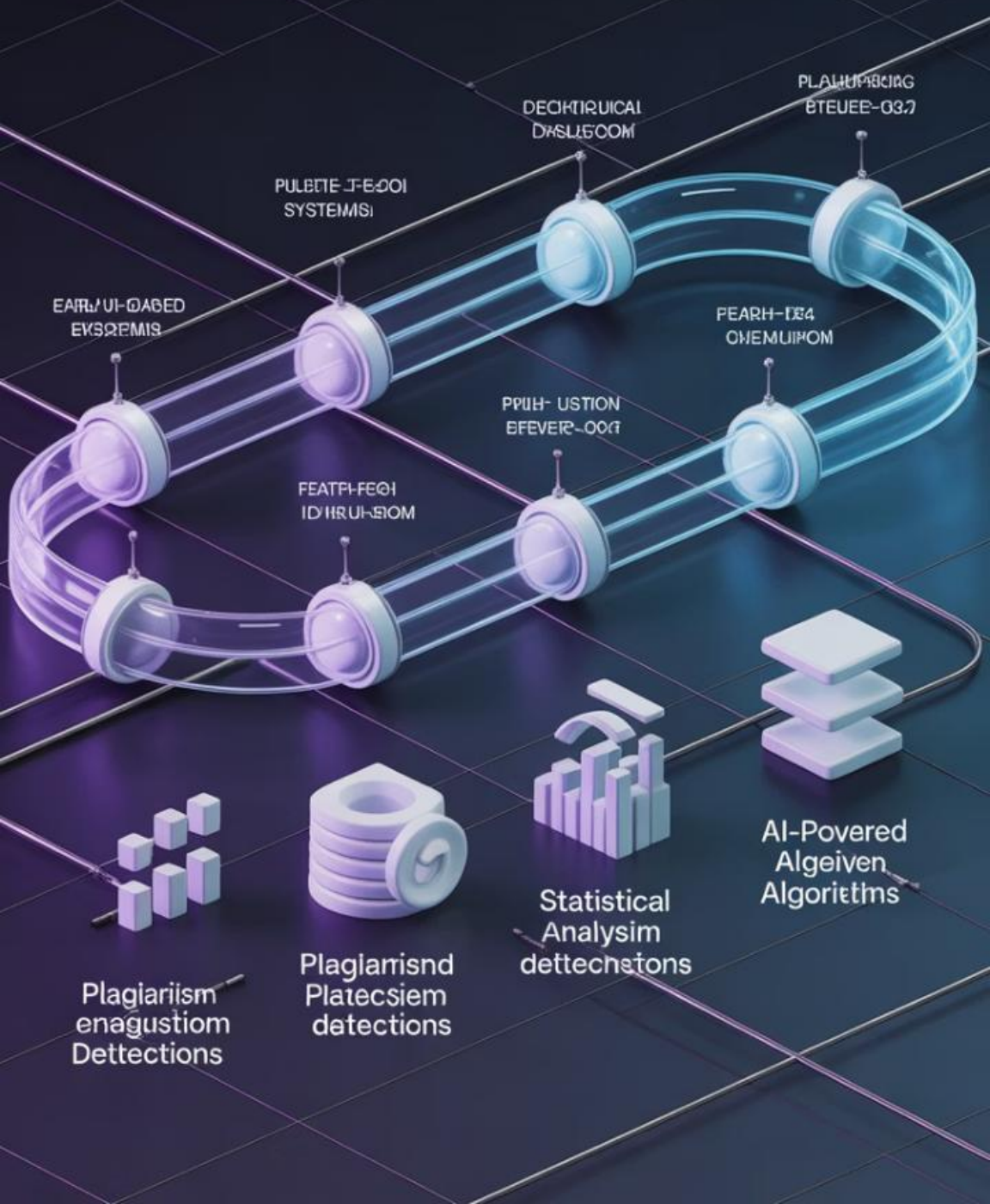
генеральный директор АО «Системы управления научными исследованиями»,
президент АО «Антиплагиат», к.ф.-м.н.

Андрей Валериевич Грабовой

доцент кафедры «Интеллектуальных систем» Московского физико-
технического института, руководитель отдела исследований АО «Антиплагиат»,
к.ф.-м.н.



Evolution of Plagiarism Detection



Содержание доклада

Прошлое

Информационный поиск на основе алгоритм шинглов

1

Настоящее

Использование новых модулей на основе ИИ

2

Будущее

ИИ – основа системы

3



Прошлое. Проблема “copy&paste”

- Использование чужих текстов стало повсеместной практикой
- Появилось большое количество ресурсов, предназначенных для распространения «готовых текстов» квалификационных работ (сайты рефератов)
- Появилось большое количество предложений по подготовке текстов «на заказ» – в основном компилятивный подход
- Традиционные средства контроля – изучение текста преподавателем, рецензентом, экспертом – не в состоянии обеспечить сопоставление полного текста документа с огромным количеством потенциальных источников. Не обеспечивается полнота охвата. Большие затраты времени
- На момент создания системы Антиплагиат в 2005 году, ни одна система не проверяла документы на русском языке

Ю. И. Журавлев, К. В. Рудаков, А. С. Инякин, А. А. Кирсанов, А. В. Лисица, Г. В. Никитов, Н. В. Песков, М.Ю. Романов, Ю. В. Чехович, Р. И. Яминов «Система распознавания интеллектуальных заимствований «Антиплагиат» // Математические методы распознавания образов. – 2005. – Т. 12, № 1. – С. 329-332.

Типы систем поиска плагиата

External plagiarism detection

Поиск совпадений с внешними источниками

Требует обширной базы данных

Сравнение текста с миллиардами документов

Intrinsic plagiarism detection

Анализ стилистических особенностей текста

Поиск аномалий в рамках одного документа

Не требует внешней базы источников

Алгоритм шинглов

Извлечение и предобработка текста

Удаление лишних символов, токенизация, удаление стоп-слов, лемматизация

Шинлирование

Формирование n-грамм (шинглов) из последовательных слов

Хеширование

Преобразование шинглов в числовые значения

Поиск кандидатов

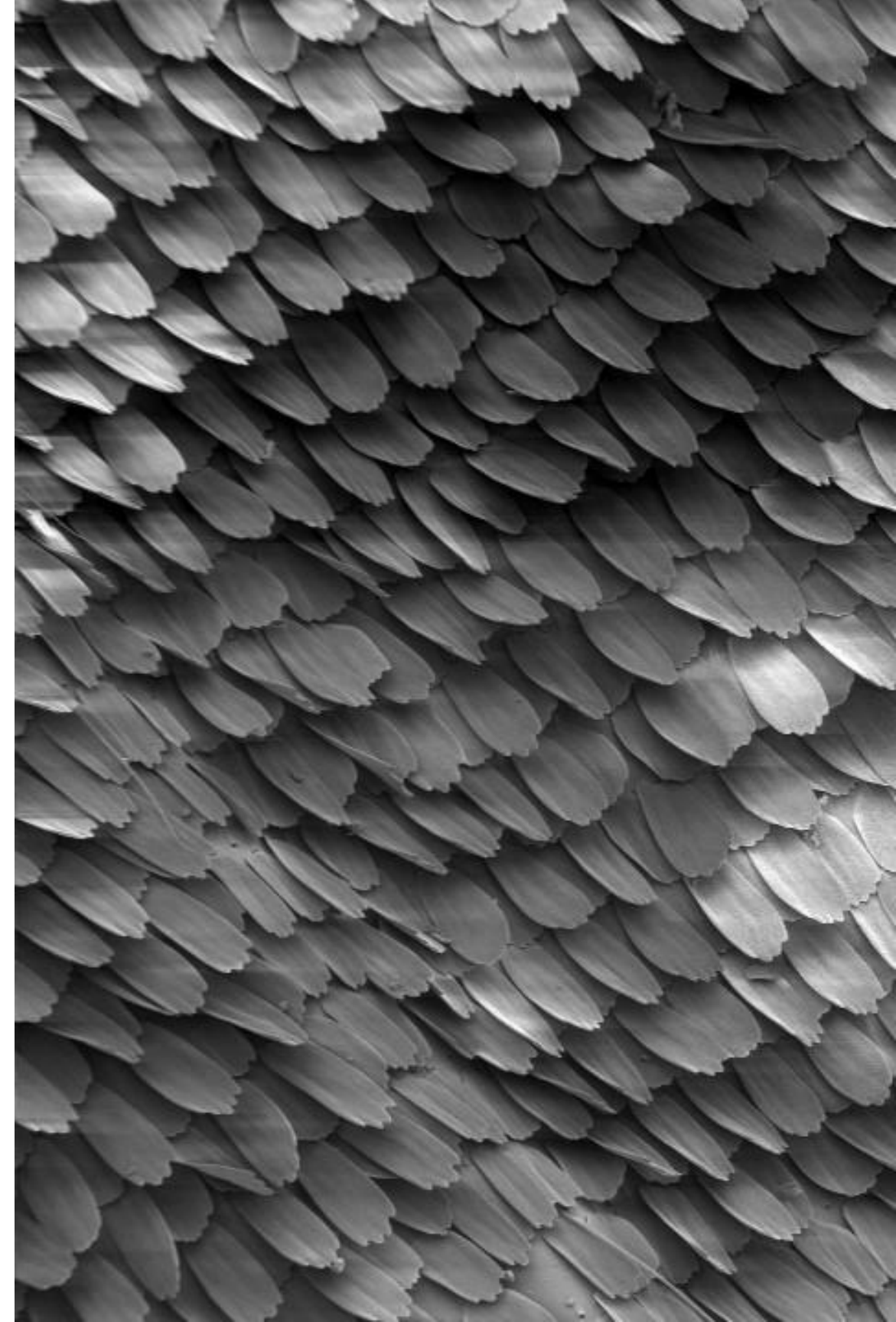
Поиск по индексу, формирование списка документов-кандидатов

Редуцирование множества кандидатов

Разностный алгоритм

Сопоставление кандидатов с запросом

Формирование отчета



Алгоритм шинглов

Указ Президента РФ от 7 мая 2012 г. N 596 "О долгосрочной государственной экономической политике"

Filter & Lemmatize

указ президент май долгосрочный государственный экономический политика

Hash

2003895740 3234756987 1887689765 3228437569 4013276519 1984765329 2987264359

Hash

4589793587629514651
9988276452874276513
7452872658426458245

Sort

4589793587629514651 7452872658426458245 9988276452874276513 ...

Ивахненко Андрей Так устроен поиск заимствований в Антиплагиате: <https://habr.com/ru/company/antiplagiat/blog/429634/>

Ботов Павел Триллион маленьких шинглов: <https://habr.com/ru/company/antiplagiat/blog/445952/>

Производительность и эффективность

Объем хранилища источников

$\sim 10^9$

документов в индексе

Пиковая производительность

до 500К

запросов в сутки

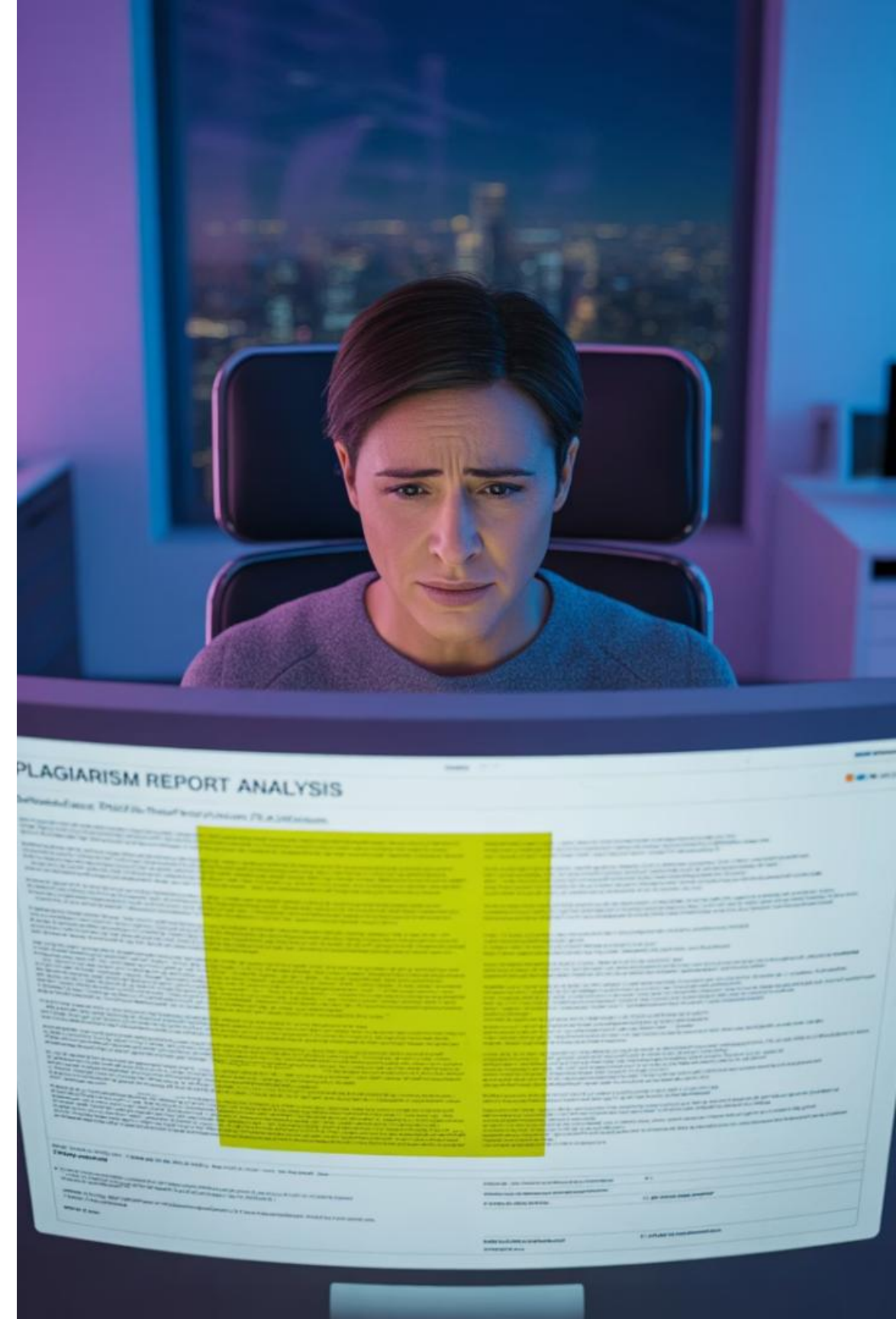


Ключевые проблемы

- Нет четких критериев плагиата (неправомерных заимствований)
- Размер заимствованного фрагмента не помогает принять решение
- Требуется интерпретировать каждый отчет, чтобы принять решение
- Используется «процентный подход» к оценке документа

Чехович Ю. В. КАК НАЙТИ ПЛОЩАДЬ ЛЕНИНА, Или размышления о практике выполнения приказа № 636, проценте оригинальности, некорректных и неправомерных заимствованиях и законе Гудхарта // Университетская книга. – 2018. – № 2. – С. 72-73.

Ю. Чехович, О. Беленькая Антиплагиат и ВКР: как не превратить проверку в фарс // Университетская книга. – 2018. – № 7. – С. 82-83.





Настоящее. ИИ-алгоритмы расширяют возможности

Появились новые вызовы, соответствовать которым стало возможным путем развития и использования новых технологий на базе машинного обучения и искусственного интеллекта

EVALUATION SYSTEM ACADEMIC INTEGRITY

Interactive Dashboard



Почему важен тип документа?

Различные подходы к оценке

Уникальные требования к каждому типу работ

Разные пороги

Допустимый процент заимствований варьируется

Исследовательская статья

Обзорная статья

Заметка

Тезисы доклада

Полный текст доклада

Курсовая работа

Дипломная работа

Магистерская диссертация

Кандидатская диссертация

Докторская диссертация

Монография

Учебник

Методические рекомендации

Заявка на грант

Отчет по проекту

Академические документы – это не только текст

Текст

Данные

Формулы

Изображения

Код программ

Ссылки

Цитаты

Структура

Journal of Informetrics 16 (2023) 101246

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/foi

Analysis of duplicated publications in Russian journals ☆☆☆☆

Yury V. Chekhovich^{1,*}, Andrey V. Khazov

Antiplagiat Company, Moscow, Russian Federation

ARTICLE INFO

Keywords:
Text recycling
Duplicate publication
Scientific ethics
Plagiarism detection

ABSTRACT

The article presents a study of publication ethics violations in Russian language scientific publications related to the duplicated publication. The aim of the study is to assess the frequency of the above-noted violations in the data of eLIBRARY.RU, the largest aggregator of full texts of Russian-language scientific publications. For the purposes of the study, we used the tools of the "Antiplagiat" plagiarism detection system. Out of the almost 12 million full-text publications on the eLIBRARY.RU platform, more than 3.8 million scientific articles were selected for analysis. The study identified 70 406 cases of duplication of publications. In each of the detected cases, the same – or significantly similar – texts were published two or more (up to 73) times. An analysis of the most significant cases by number of publications is presented along with a detailed discussion of examples of the identified violations. A significant increase in the number and proportion of duplicated publications was identified in the period from 2014 to 2017. Conclusions are presented concerning shortcomings in editorial processes that allow duplication of publications along with the factual impossibility of detecting duplication in cases of simultaneous submission of the manuscript to different journals, the insufficient use by journal editors of means for detecting inappropriate borrowings and the need to retract a significant number of articles in connection with the identified violations.

1. Introduction

In recent years, the phenomenon of text recycling – i.e., the use (ethical or not) by authors of portions of text from their previously published articles – has been the subject of many works on academic ethics (Moskovitz, 2019; Roig, 2015). Despite some terminological differences, two main types of text recycling can be distinguished. The first of these is duplicate publication, which is defined as the publication of an article that is identical or overlaps substantially with an article already published elsewhere, with or without acknowledgement (Benos et al., 2005). In other words, the same text is republished by the same authors (or at least one of the co-authors). In such cases, the title, abstract, and keywords may be deliberately altered. The earliest identified mention of the phenomenon of duplicate publication (Watanakunakorn, 1975) refers to publications in 1971 (Lee & Kerstein, 1971) and 1973 (Kerstein & Lee, 1973). The second type of text recycling is redundant publication, where authors reuse a significant quantity of text from a previously published work in a new article. In this case, the previously published text can be supplemented with additional text and results.

^{*} Declarations Funding: The research was funded by Antiplagiat Company

^{**} Conflicts of interest/Competing interests: Authors are the employees of Antiplagiat company

^{***} Availability of data and material: All the results obtained are in the public repository: <http://dx.doi.org/10.17632/dy2smg277.2>

^{****} Code availability: For the study, we used a commercial software Antiplagiat system (www.antiplagiat.ru).

[†] Corresponding author at: Vostochnaya str. 11 – 1 – 58, Moscow 115280, Russian Federation.
E-mail addresses: chekhovich@ap-team.ru (Y.V. Chekhovich), khazov@ap-team.ru (A.V. Khazov).

¹ ORCID: 0000-0002-5204-5484.

<https://doi.org/10.1016/j.foi.2021.101246>

Received 15 June 2021; Received in revised form 29 November 2021; Accepted 22 December 2021

Available online 6 January 2022

1751-1577/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



Извлечение метаданных

Название работы

Идентификация заголовка документа

Авторы

Распознавание имен создателей

Дата публикации

Определение времени создания

Библиографические записи

Извлечение списка источников



Извлечение метаданных

Используемые методы:

- Регулярные выражения для структурированных титульных листов
- Небольшие рекуррентные сети, для решения задач классификации

Плюсы «легких» моделей — работают быстро и не потребляют много ресурсов

Современные подходы на основе методов выделения именованных сущностей и других смысловых элементов:

- Дообучения современных transformer моделей — повышение полноты поиска на 10% (70->80)
- Снижение скорости обработки документов, высокие требования к вычислительным ресурсам

Определение структуры документа

Различный подход к оценке

В разных частях текста допустимы разные заимствования

- Введение может содержать больше цитат
- Методы часто стандартизированы
- Результаты должны быть уникальными
- Библиография содержит стандартизованные ссылки на публикации, но совпадение большого количества позиций подозрительно

Типы структурных элементов

- Титульный лист*
- Содержание*
- Введение*
- Методы*
- Основная часть*
- Выводы библиография*
- Приложения*



Определение структуры документа

- Используется не только текст работы, а также и информации о форматировании документа. Это позволяет значительно увеличить качество работы, без значимого увеличения сложности модели.
- Важна априорная информация о последовательности глав в документе. Но есть случаи, когда она нарушается, поэтому за нарушения последовательности стоит штрафовать, но не значительно.
- Современные модели на базе трансформеров не дают значимого прироста качества для этого класса задач на наших данных по сравнению с используемыми более простыми моделями

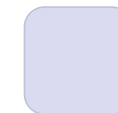
Ilia Kopanichuk, Artem Chashchin, Inga Ochneva, Andrey Grabovoy, Aleksandr Ogaltsov, Aleksandr Kildyakov, Yury Chekhovich "Structure Extractor: Multilingual Extraction of Sections from Scientific Document," 2025 37th Conference of Open Innovations Association (FRUCT), Narvik, Norway, 2025, pp. 122-128, doi: 10.23919/FRUCT65909.2025.11008070.

Причины распространения переводного плагиата



Развитие машинного перевода

Доступность качественных систем перевода



Рост международных баз данных

Доступ к публикациям на разных языках



Обход классических систем

Трудности обнаружения переведенного текста

linguaflow
Unlock global knowledge



Кросс-языковой поиск заимствований

1

- Машинный перевод документа

2

- Кластеризация слов по синонимии

3

- Шинглирование по номерам кластеров

4

- Поиск документов-кандидатов

5

- Точное сопоставление и определение границ фрагментов

- O. Bakhteev, R. Kuznetsova, A. Romanov and A. Khritankov, "A monolingual approach to detection of text reuse in Russian-English collection," 2015 Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference (AINL-ISMW FRUCT), 2015, pp. 3-10
- Avetisyan, K., Gritsay, G. & Grabovoy, A. " Cross-Lingual Plagiarism Detection: Two Are Better Than One " Program Comput Soft 49, 346–354 (2023).
- Oleg Bakhteev Yury Chekhovich, Andrey Grabovoy et al. (2022). "Cross-Language Plagiarism Detection: A Case Study of European Languages Academic Works" In: Academic Integrity: Broadening Practices, Technologies, and the Role of Students. Ethics and Integrity in Educational Contexts, vol 4. Springer, Cham.

Обнаружение парафраза

Поиск парафраза является частным случаем поиска кросс-языковых заимствований при условии совпадении языков документа-запроса и проиндексированного документа

2

- Кластеризация слов по синонимии

3

- Шинглирование по номерам кластеров

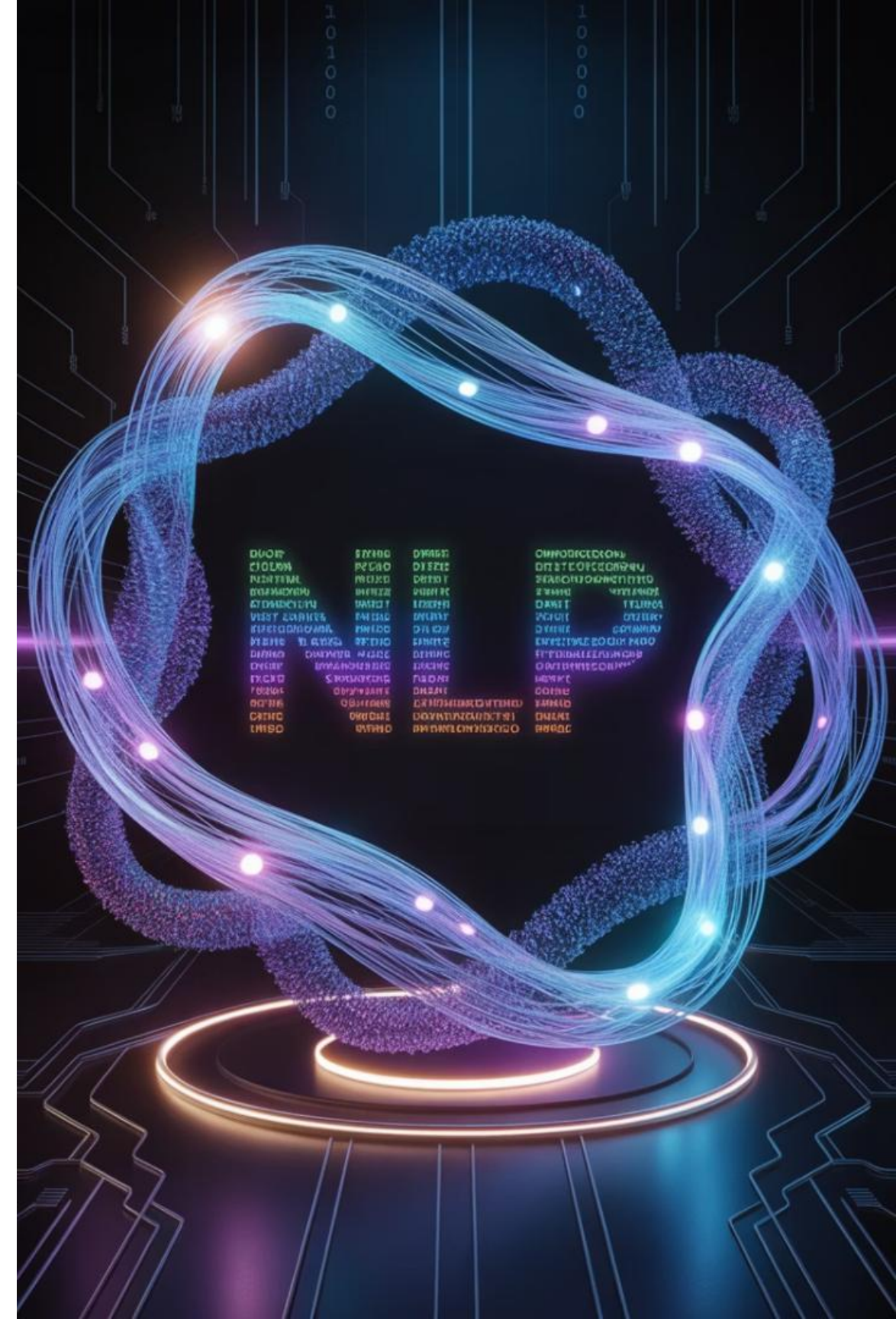
4

- Поиск документов-кандидатов

5

- Точное сопоставление и определение границ фрагментов

Юрий Чехович «Трое в лодке, нищета и собаки», или как Антиплагиат ищет парафраз / <https://habr.com/ru/companies/antiplagiat/articles/422941/>



Поиск заимствованных изображений

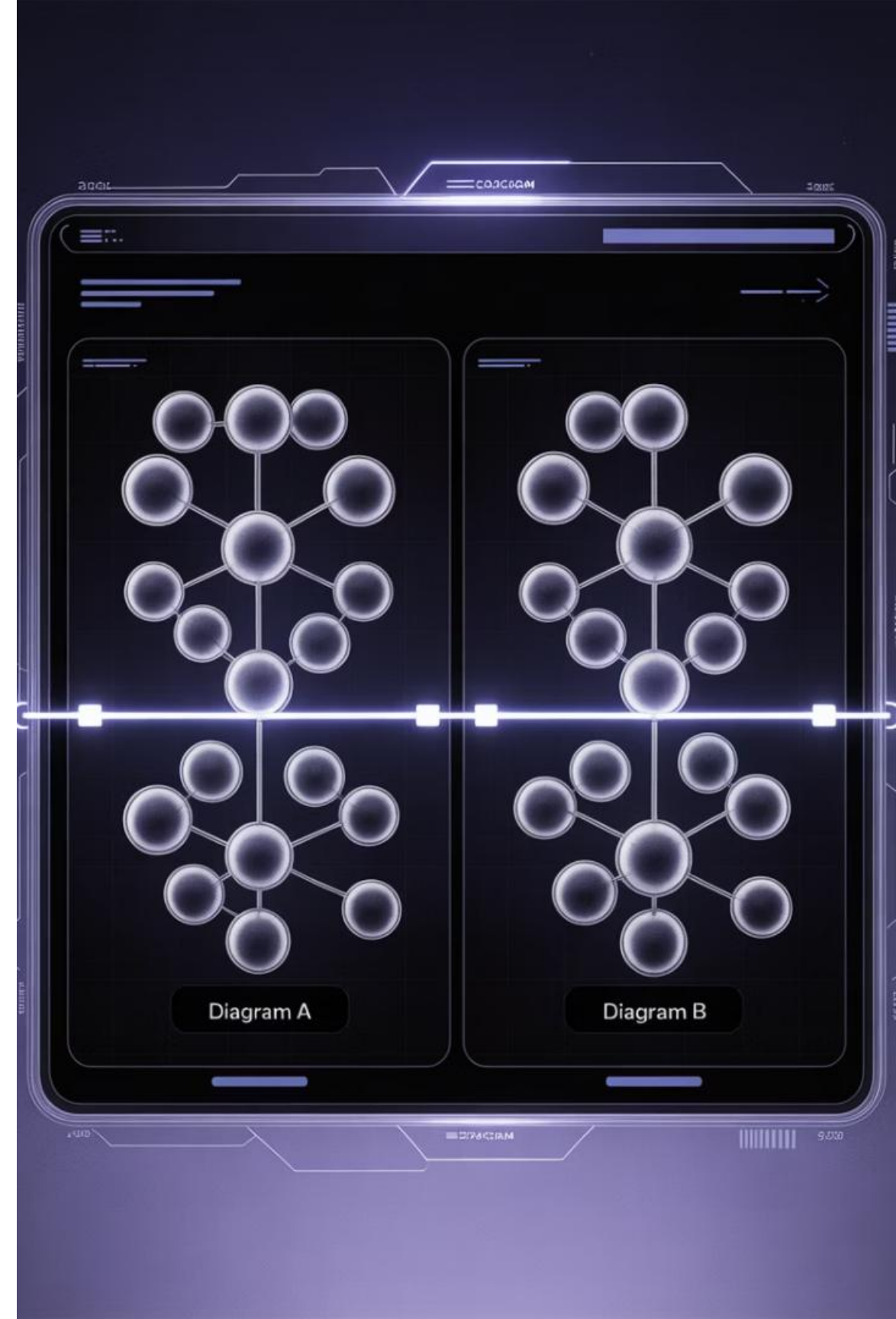
Существует множество научных работ, значительная часть содержания которых заключена в изображениях

Распространенные изменения:

- Слабое кадрирование (до 5% от оригинального размера)
- Умеренное изменение пропорций (не более 25% от оригинального размера)
- Перевод в градации серого
- Размытие

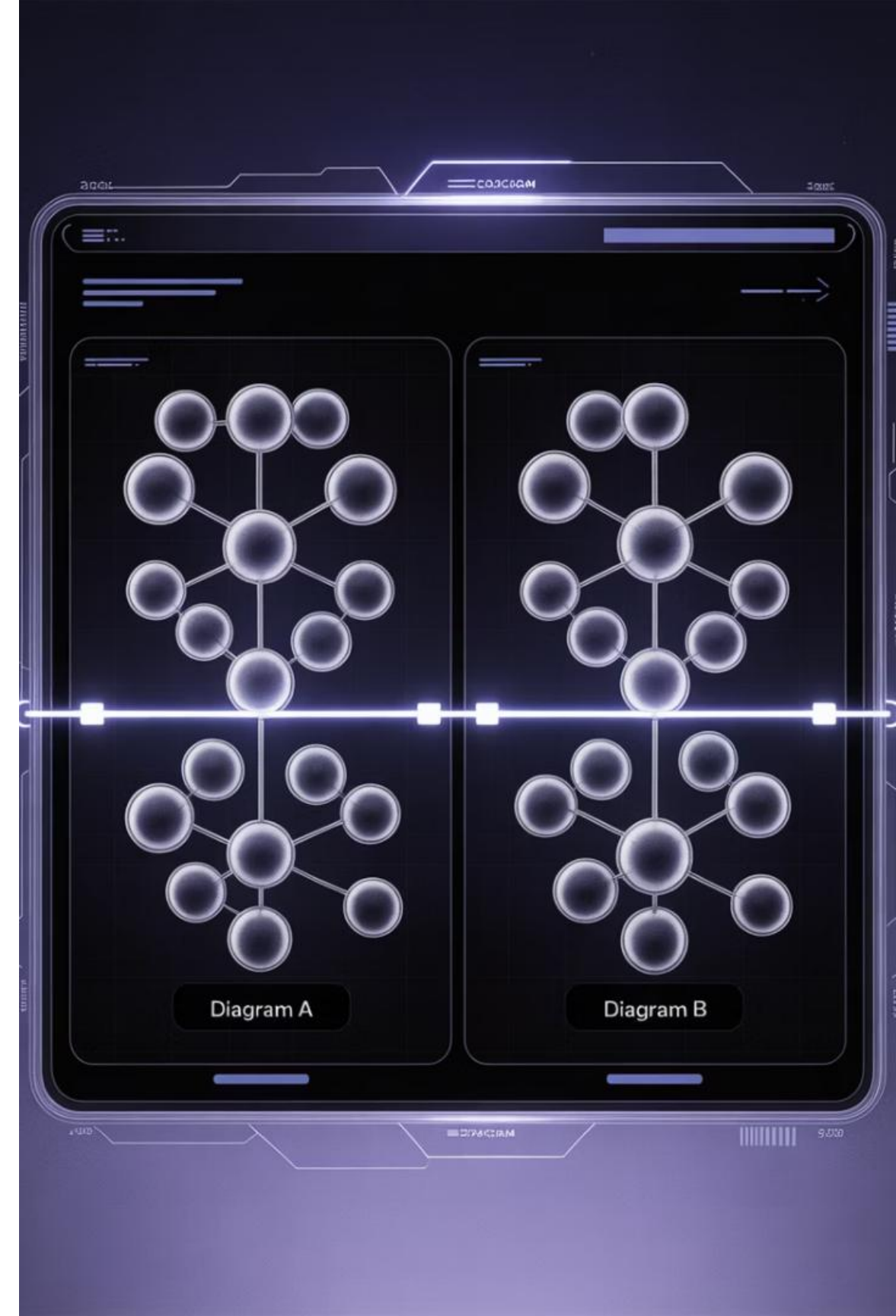
Редкие изменения:

- Значительное кадрирование (до 15% от оригинального размера)
- Значительное изменение пропорций (не более 50% от оригинального размера)
- Поворот на 90, 180, 270 градусов

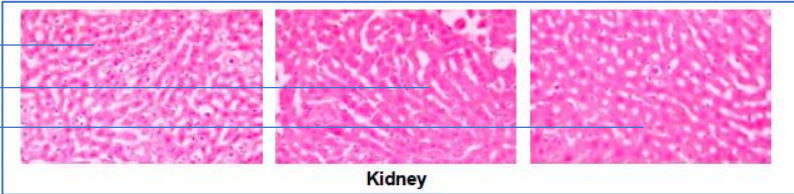
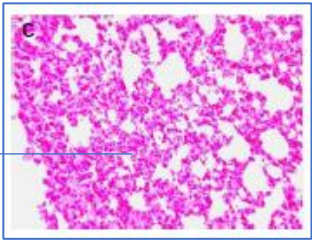
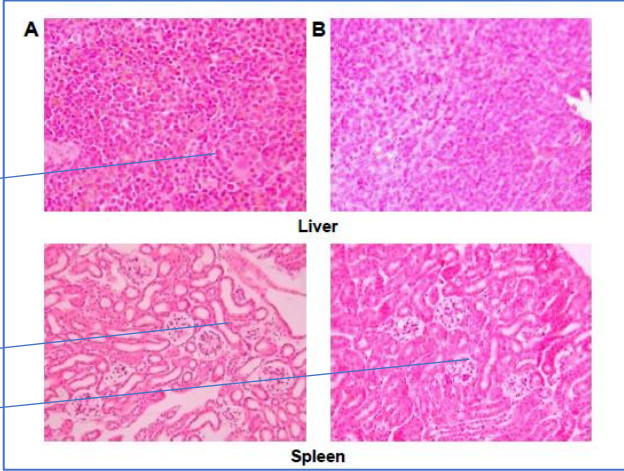
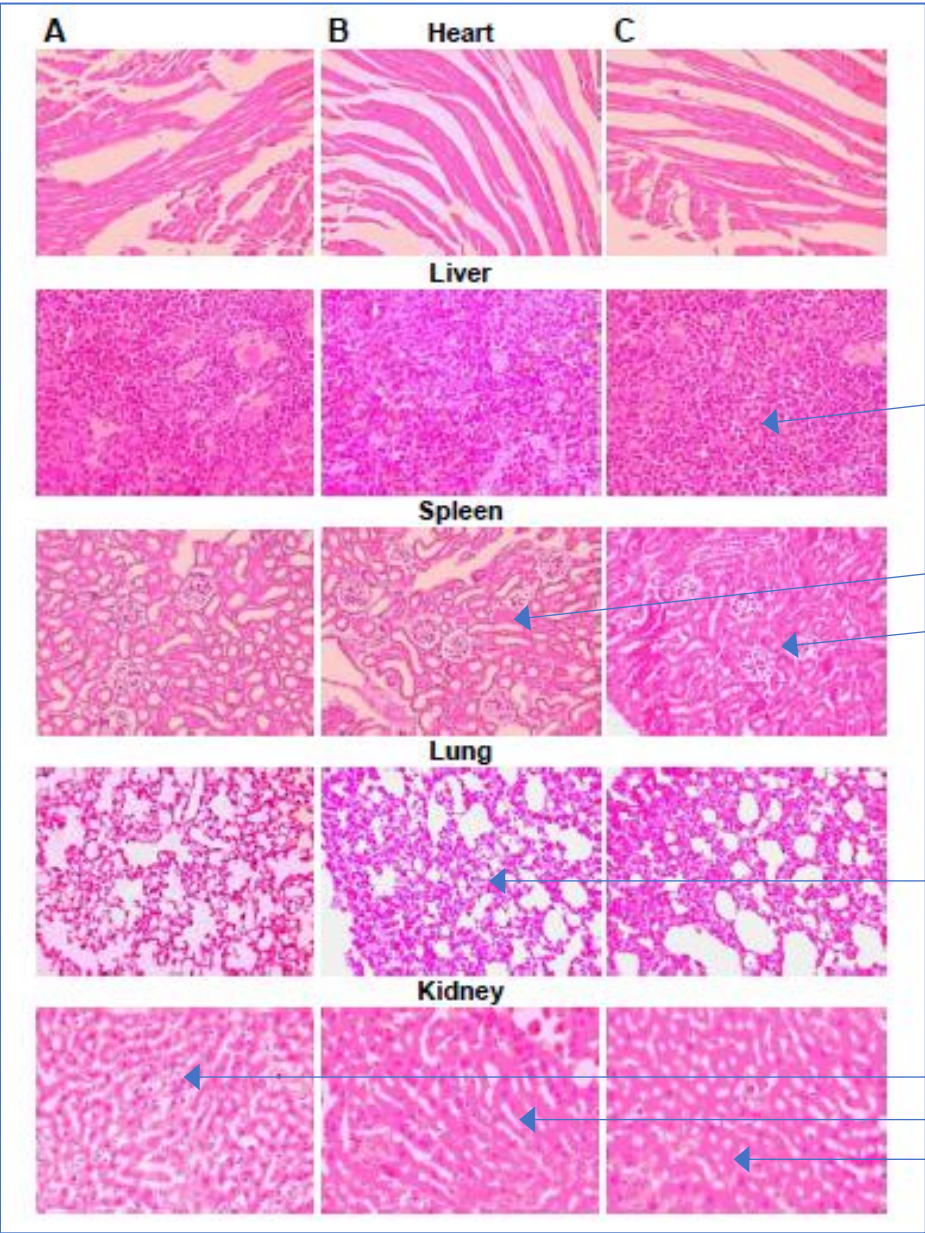


Поиск заимствованных изображений

- Построение векторного представления изображения. Использовалась архитектура Deep Ranking*
 - Поиск кандидатов. Использовалась векторная база данных FAISS
 - Точное сравнение кандидатов с запросом. Использовалась архитектура типа Siamese Neural Network
-
- М.С. Каприелова, К.В. Варламова, И.О. Потяшин, Ю.В. Чехович. Поиск почти дубликатов изображений рукописных текстов для высоконагруженных сервисов // Известия Российской академии наук. Теория и системы управления. - 2024. - №4. - С. 129-139.
 - M. Kaprielova, A. Grabovoy, I. Potyashin, K. Varlamova, Y. Chekhovich. Image Plagiarism Detection Pipeline for Vast Databases // 2024 35th Conference of Open Innovations Association (FRUCT). – IEEE, 2024. – С. 328-335.



Поиск заимствованных изображений. Плагиат



Li S, Wang X. In vitro and in vivo evaluation of novel NGR-modified liposomes containing brucine. *Int J Nanomedicine*. 2017;12:5797-5804
<https://doi.org/10.2147/IJN.S136378>

Li D, Gong L. Preparation of novel pirfenidone microspheres for lung-targeted delivery: in vitro and in vivo study. *Drug Des Devel Ther*. 2016;10:2815-2821
<https://doi.org/10.2147/DDDT.S113670>

Chen J, Jiang H, Wu Y, Li Y, Gao Y. A novel glycyrrhetic acid-modified oxaliplatin liposome for liver-targeting and in vitro/vivo evaluation. *Drug Des Devel Ther*. 2015;9:2265-2275
<https://doi.org/10.2147/DDDT.S81722>

Zhou X, Tao H, Shi KH. Development of a nanoliposomal formulation of erlotinib for lung cancer and in vitro/in vivo antitumoral evaluation. *Drug Des Devel Ther*. 2018;12:1-8 <https://doi.org/10.2147/DDDT.S146925>

Поиск заимствованных изображений. Фальсификация

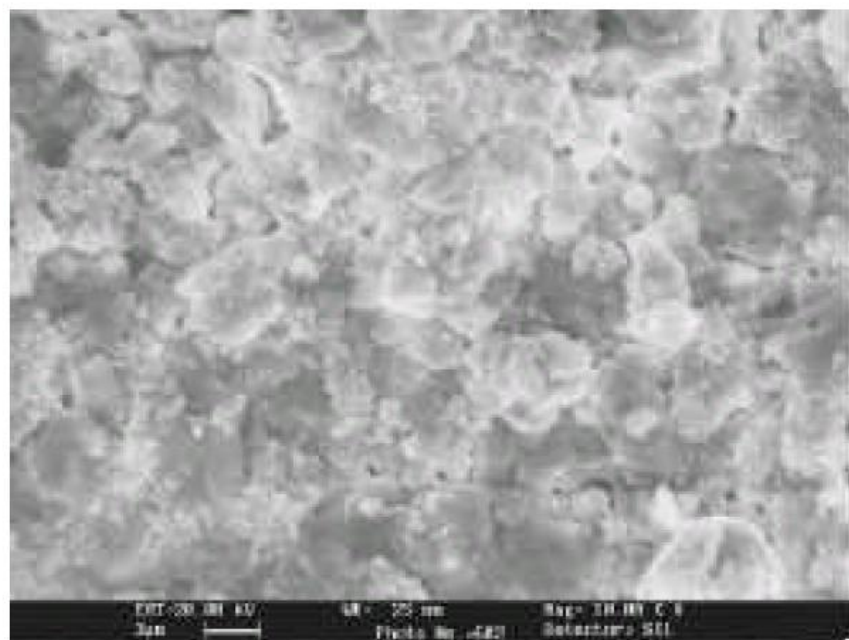


Fig. 5c. SEM image of MgSA6 composite.

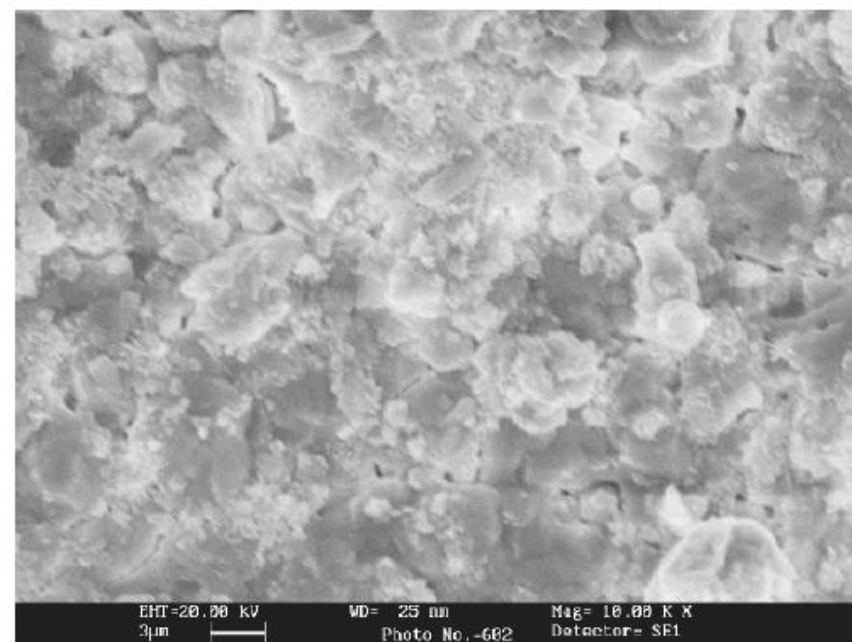
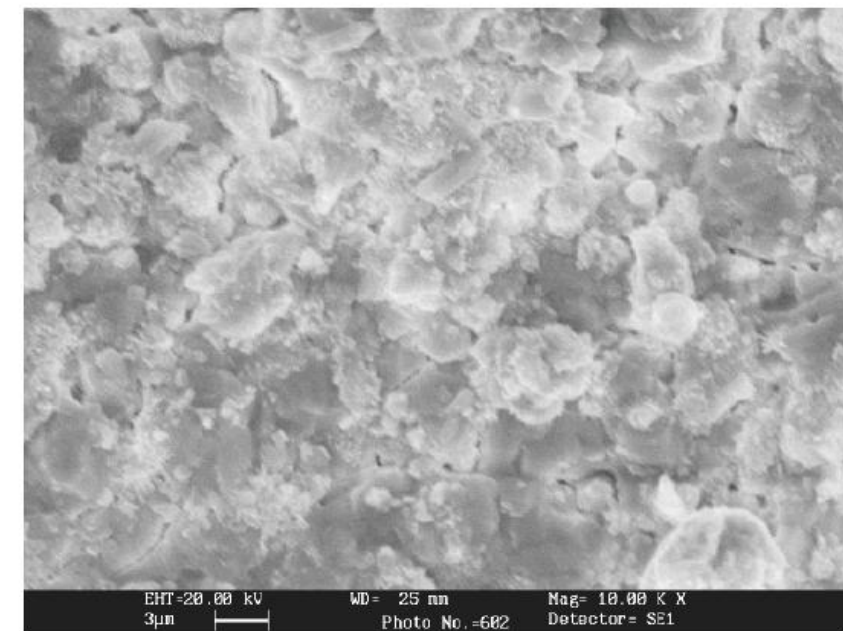


Fig. 5c. SEM image of ZnSA6 composite.



(c)

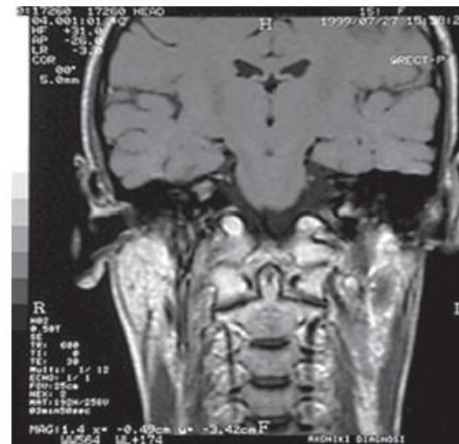
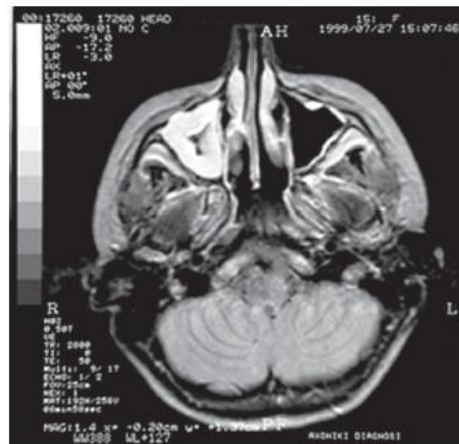
Fig. 3. SEM image of a) BaSA1, b) BaSA5 and c) SA composite.

J. JUDITH VIJAYA, L. JOHN KENNEDY, G. SEKARAN, K.S. NAGARAJA
Methanol Sensing Behavior of Strontium(II) Added MgAl₂O₄ Composites Through Solid-State Electrical Conductivity Measurements
Sensors & Transducers Journal, Vol.74, Issue 12, December 2006, pp.864-873

J. JUDITH VIJAYA, L. JOHN KENNEDY, G. SEKARAN, K.S. NAGARAJA
Synthesis, Characterization and Acetone Sensing Properties of Novel Strontium(II)-added ZnAl₂O₄ Composites
Sensors & Transducers Journal, Vol.76, Issue 2, February 2007, pp.1008-1017

B. Jeyaraj, L. John Kennedy, G. Sekaran and J. Judith Vijaya
Benzene and Toluene Vapor Sensing Properties of Sr(II)-added Barium Aluminate Spinel Composites
Sensors & Transducers Journal, Vol. 96, Issue 9, September 2008, pp. 68-80

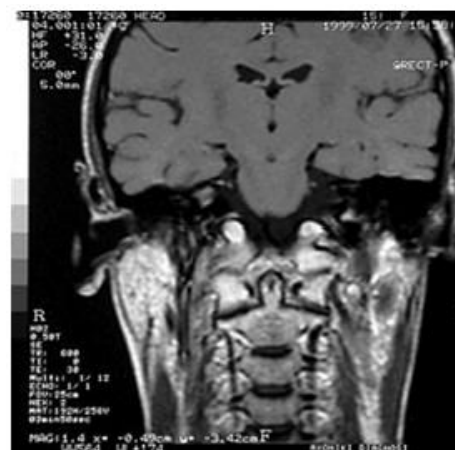
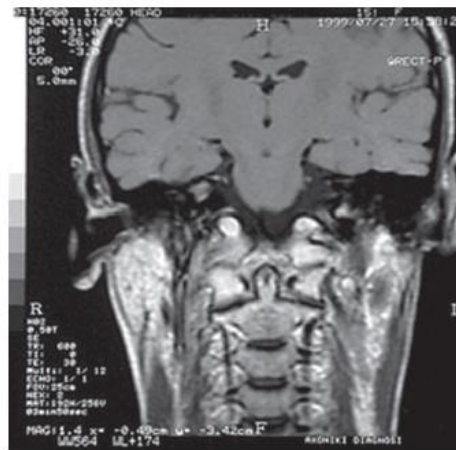
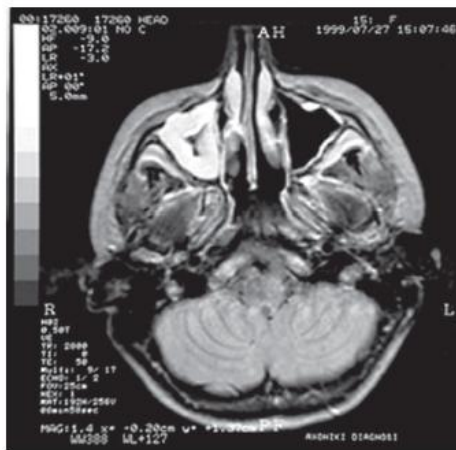
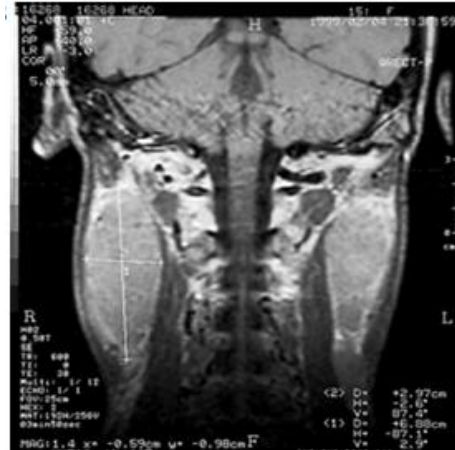
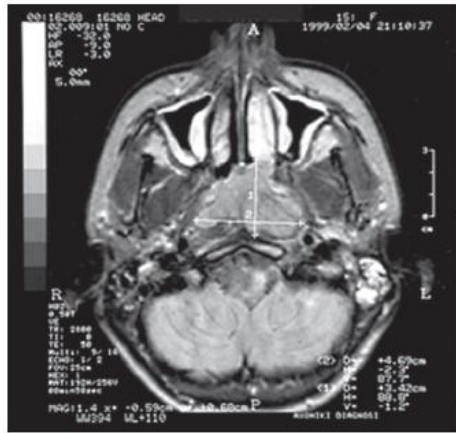
Поиск заимствованных изображений. Фальсификация



Nikolaos Eleftheriadis, Christos Papaloukas, Damianos Eleftheriadis, Apostolos Hatzitolios, Ioulia Ioannidou-Marathiotou & Kiki Pistevou-Gompaki Long-term radiotherapy related complications in children with head and neck cancer: Another era for pediatric oncologic pathology // International Journal of General Medicine 2009:2 63–66

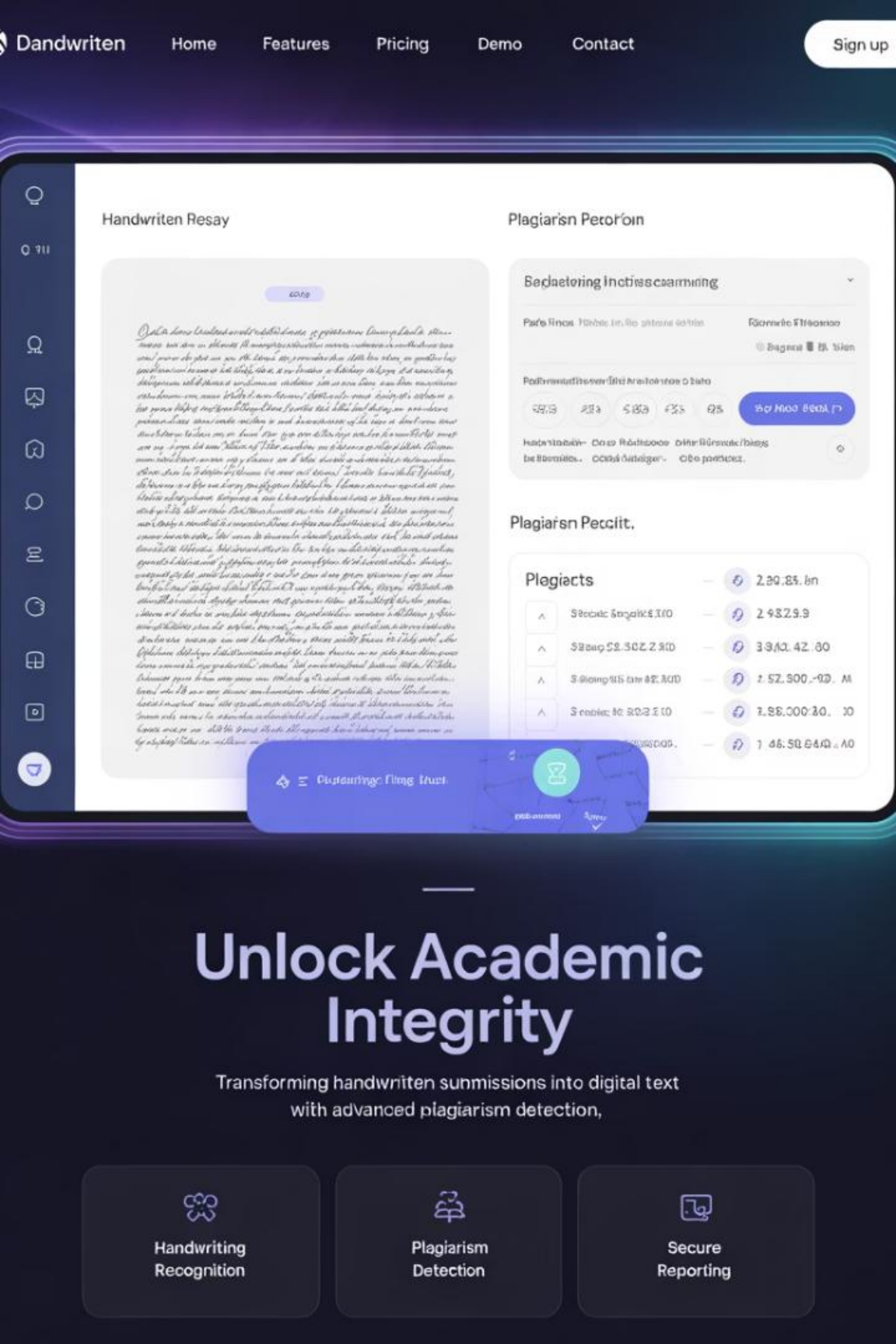
Ioulia Ioannidou-Marathiotou, Kyriaki Pistevou-Gompaki, Nikolaos Eleftheriadis & Christos Papaloukas Long term chemoradiotherapy-related dental and skeletal complications in a young female with nasopharyngeal carcinoma // International Journal of General Medicine 2010:3 187–196

Поиск заимствованных изображений. Фальсификация



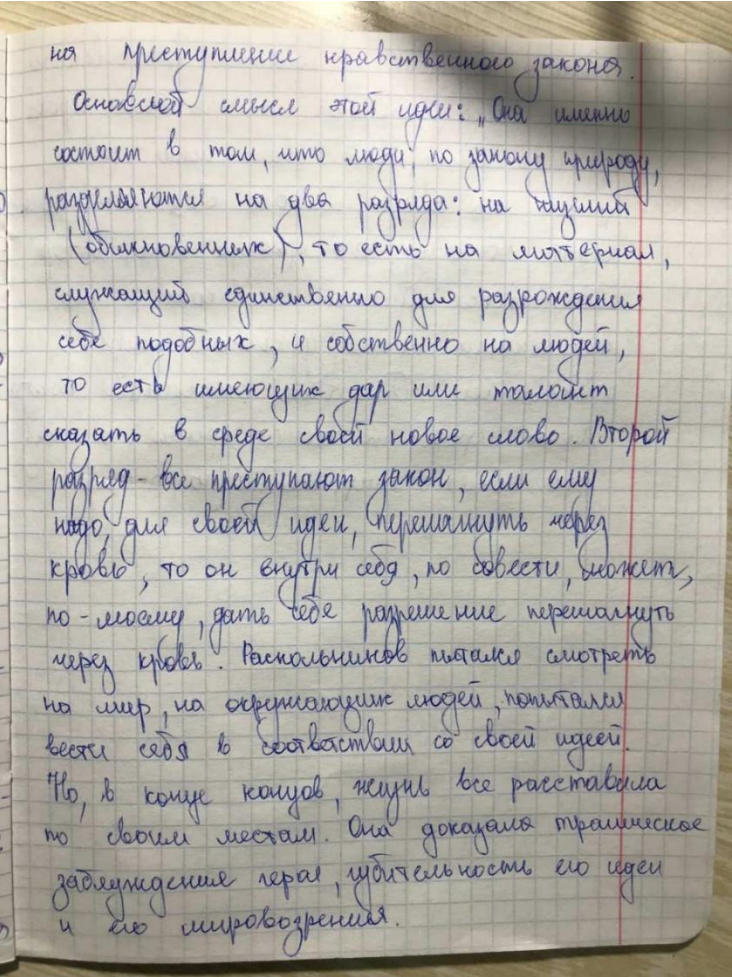
Abstract: ...We report on **two male children (8 and 14 years old)** with head and neck cancer, who were successfully treated with chemoradiotherapy and presented with growth deficiency of middle face and mandible hypoplasia, eight years and one year later, respectively...

Abstract: We describe the long-term complications six years after chemoradiotherapy in **a 20-year old woman** with nasopharyngeal carcinoma. We wanted to know whether the radiation dose was constant throughout the oral cavity, and thus uniformly affecting the corresponding dental and skeletal structures...

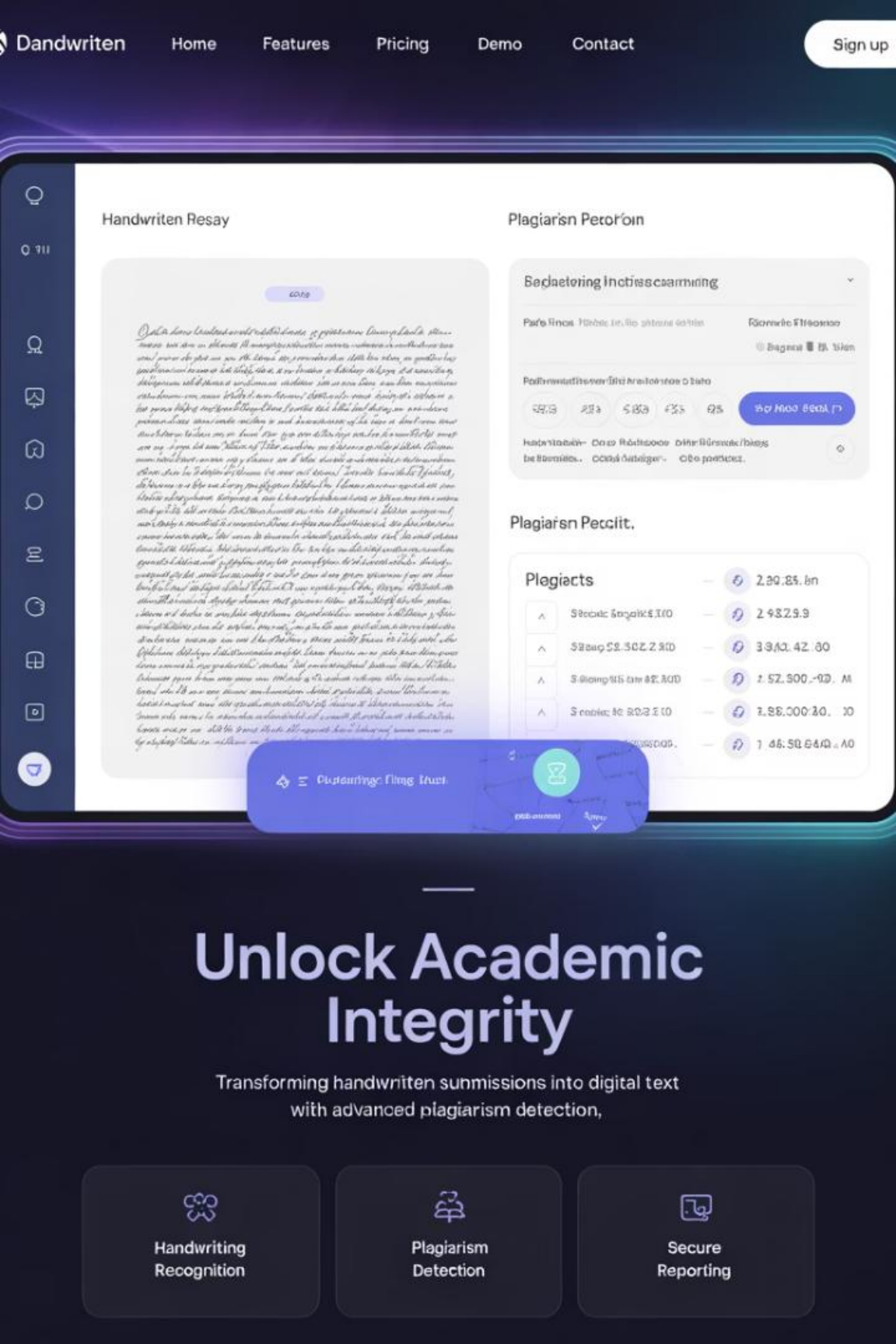


Обнаружение плагиата в рукописных работах

Рукописные сочинения школьников в соответствии с ФГОС необходимо проверять на наличие заимствований



на преступлении правственного закона. основной смысл этой идеи: она именно состоит в том, что люди, по закону природы, раздвигаются на два разряда; на куелий бымновенных), то есть на мотвериял, служащий сдинетвенно для разрождения себ подобных, и собственно на людей, го есть имеуюих дар или талонт сназать в среде своей новое слово. Второй родряд- все преступают закон, если ему нодо, для своей идеи, перешагнуть через кровь, то он внутри себя, по совести, может, по- мосу, дать себя разрешение перешалнуть череу кровь. Раснольшнов потался смотреть на мир, на очрежаюдшик людей, поньталия вести себя в соствастваи со своей идеей. Но, в конце колуов, жунь все расставала по своим метам. Она доказала трашическое заболуждения перая, убительность его иден и ко смуовозрения.



Обнаружение плагиата в рукописных работах

Предобработка

Определение ориентации
Выделение строк

Распознавание текста

Выявление хорошо распознаваемых символов
Выделение биграмм, с хорошим качеством распознавания

Поиск кандидатов

Шинглирование
Поиск по индексу

Сопоставление кандидатов с запросом

Векторизация совпадений
Оценка близости

- A. V. Grabovoy, M. S. Kaprielova, A. S. Kildyakov, I. O. Potyashin, T. B. Seyil, E. L. Finogeev & Yu. V. Chekhovich Text reuse detection in handwritten documents // Doklady Mathematics, 2024.
- Ivan Potyashin, Mariam Kaprielova, Yury Chekhovich, Alexandr Kildyakov, Temirlan Seil, Evgeny Finogeev & Andrey Grabovoy HWR200: New open access dataset of handwritten texts images in russian // Proceedings of the International Conference "Dialogue". — Vol. 2023. — 2023.

Handwritten Resay

Plagiarism Detector

Handwriting Recognition

Plagiarism Detection

Secure Reporting

Unlock Academic Integrity

Transforming handwritten submissions into digital text with advanced plagiarism detection,

Handwritten Resay

Plagiarism Detector

Handwriting Recognition

Plagiarism Detection

Secure Reporting

Unlock Academic Integrity

Transforming handwritten submissions into digital text with advanced plagiarism detection,

Обнаружение плагиата в рукописных работах

Handwritten Russian text in a notebook. At the top, there is a table with student names and dates:

Имя	Фамилия	Дата
Думалов	Арина	9
Александровна	Иванов	10
	Иванов	17

Below the table, there is a section titled "Задание 1" (Task 1) with handwritten text in Russian. The text describes a task involving a mountain and a river, and includes a list of tasks (1-5) related to the text.

Handwritten Russian text on a piece of paper. At the top, there is a table with student names and dates:

Имя	Фамилия	Дата
Думалов	Арина	9
Александровна	Иванов	10
	Иванов	17

Below the table, there is a section titled "Задание 1" (Task 1) with handwritten text in Russian. The text describes a task involving a mountain and a river, and includes a list of tasks (1-5) related to the text.

27/36



Обнаружение заимствований в программном коде



Токенизация

Разбиение кода на логические единицы



Создание сигнатуры

Формирование уникального идентификатора

2

Анализ структуры

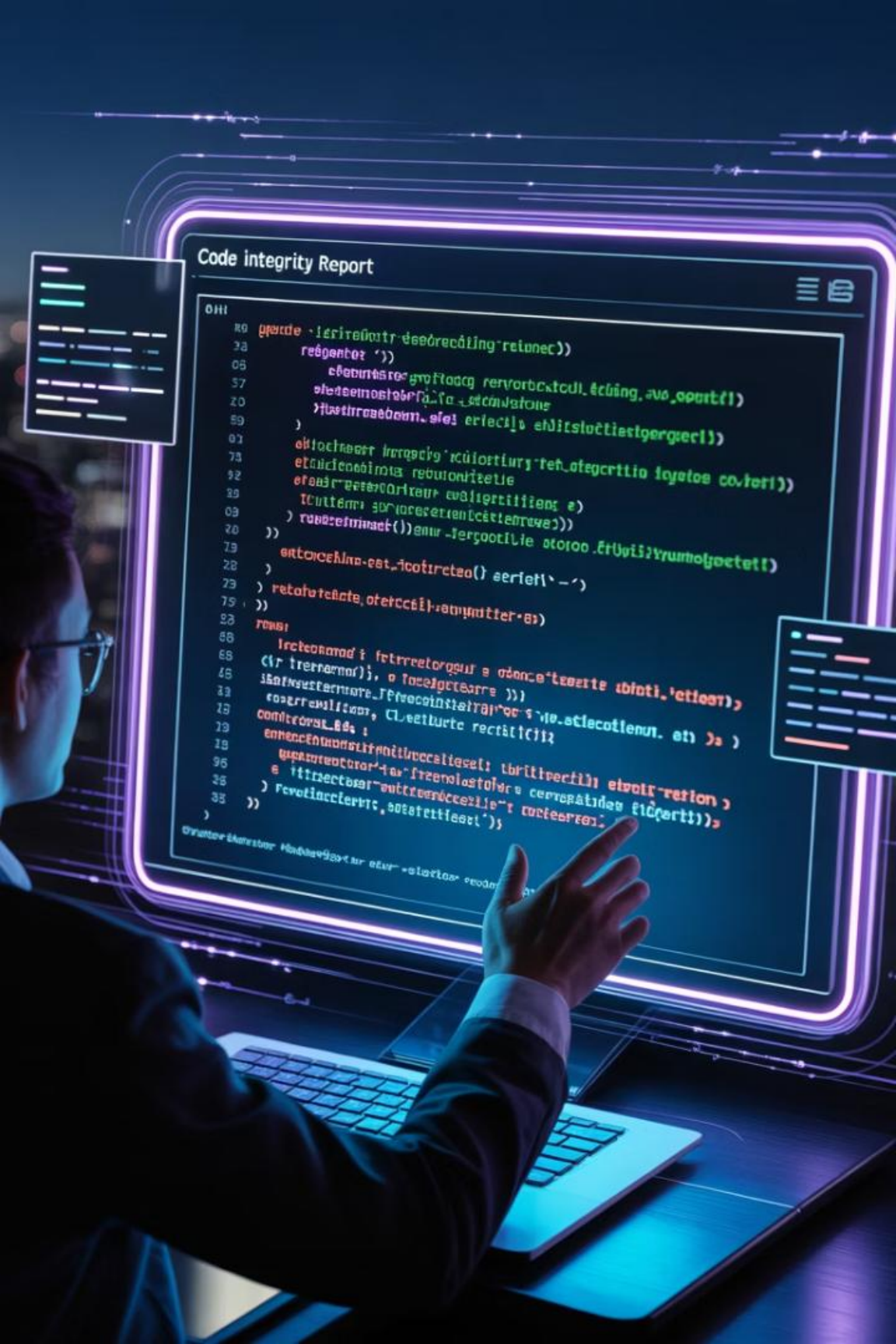
Изучение алгоритмической организации



Поиск совпадений

Выявление аналогичных фрагментов

Особенности поиска в коде



Устойчивость к переименованию

Идентификация схожести при смене имен переменных

Анализ логики

Обнаружение схожих алгоритмических решений

Учет библиотек

Исключение стандартных фрагментов

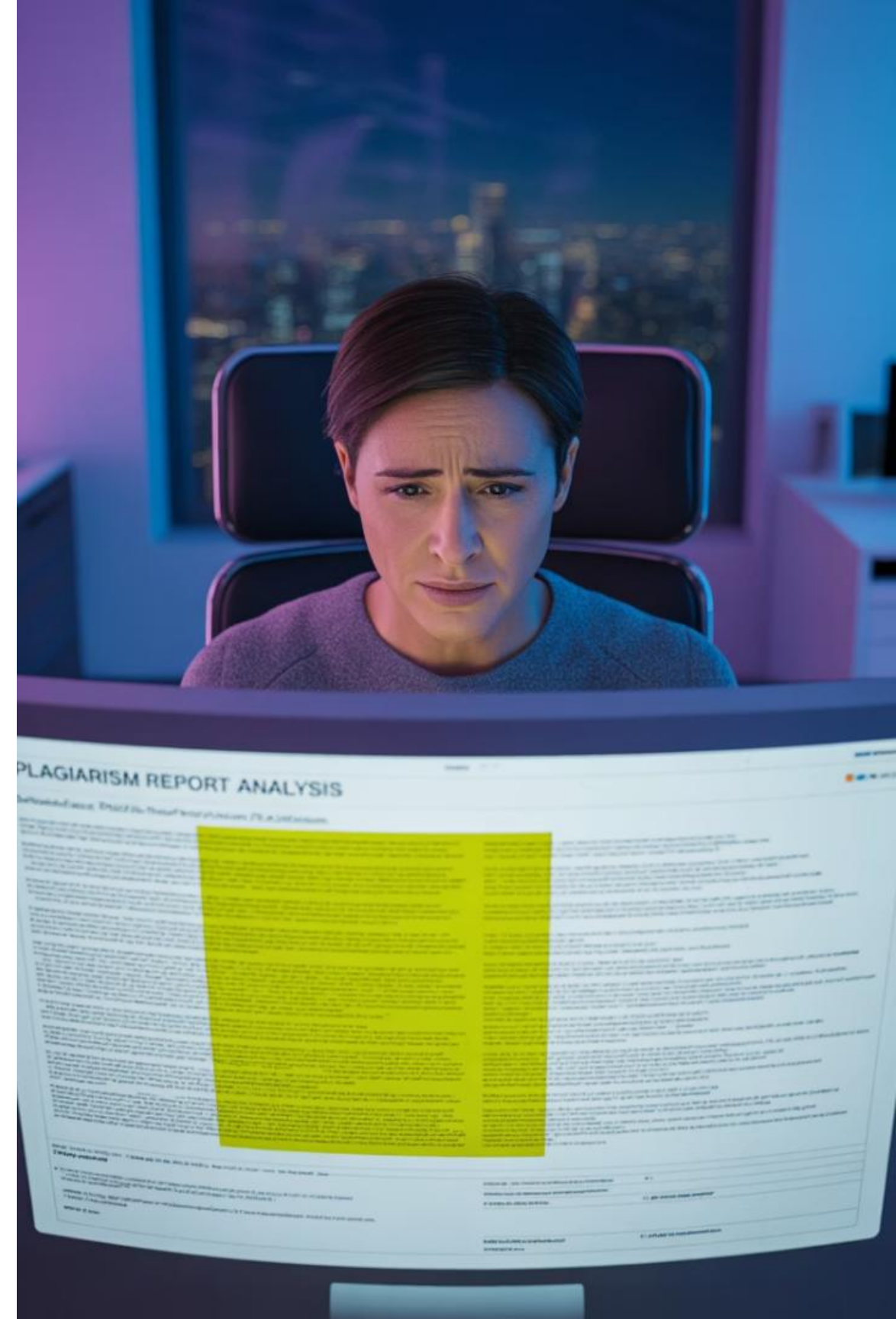
Детекция искусственного текста

- 30 ноября 2022 года OpenAI представил ChatGPT
- С 10 мая 2023 года в системе Антиплагиат работает детектор искусственного текста

- G. Gritsay, A. Grabovoy, Y. Chekhovich. (2022). "Automatic Detection of Machine Generated Texts: Need More Tokens"
- G. Gritsay, A. Grabovoy, Y. Chekhovich, A. Kildyakov (2023). "Automated Text Identification: Multilingual Transformer-based Models Approach"
- Gritsay, G.M., Grabovoy, A.V., Kildyakov, A.S. et al. Artificially Generated Text Fragments Search in Academic Documents. Dokl. Math. 108 (Supl 2), S434–S442 (2023)
- G. Gritsay, A. Grabovoy, Y. Chekhovich (2024). "Generative AI Models with Their Full Reveal"

Ключевые проблемы

- Значительно выросли затраты времени и вычислительных ресурсов на обработку одного документа
- Стал существенно сложнее анализ отчета
- Выросли требования к пользователям, которые могут проводить анализ отчета





Будущее. Интеллектуализация систем

Технологический рывок в области систем ИИ создает предпосылки для интенсивного развития систем обнаружения заимствований

Будущее систем обнаружения заимствований



1

Автоматическая интерпретация отчетов

ИИ-системы поддержки принятия решений

2

Доказательная детекция ИИ

Объяснимые алгоритмы выявления искусственного текста

3

Новые форматы

Анализ видео, аудио и презентаций

4

Диалоговые сервисы

Обучающие системы с обратной связью

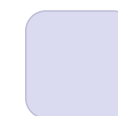


Автоматическая интерпретация отчетов



Оценка критичности

Автоматическое определение серьезности нарушений



Формирование рекомендаций

Предложения по дальнейшим действиям



Персонализация критериев

Адаптация к требованиям конкретной организации

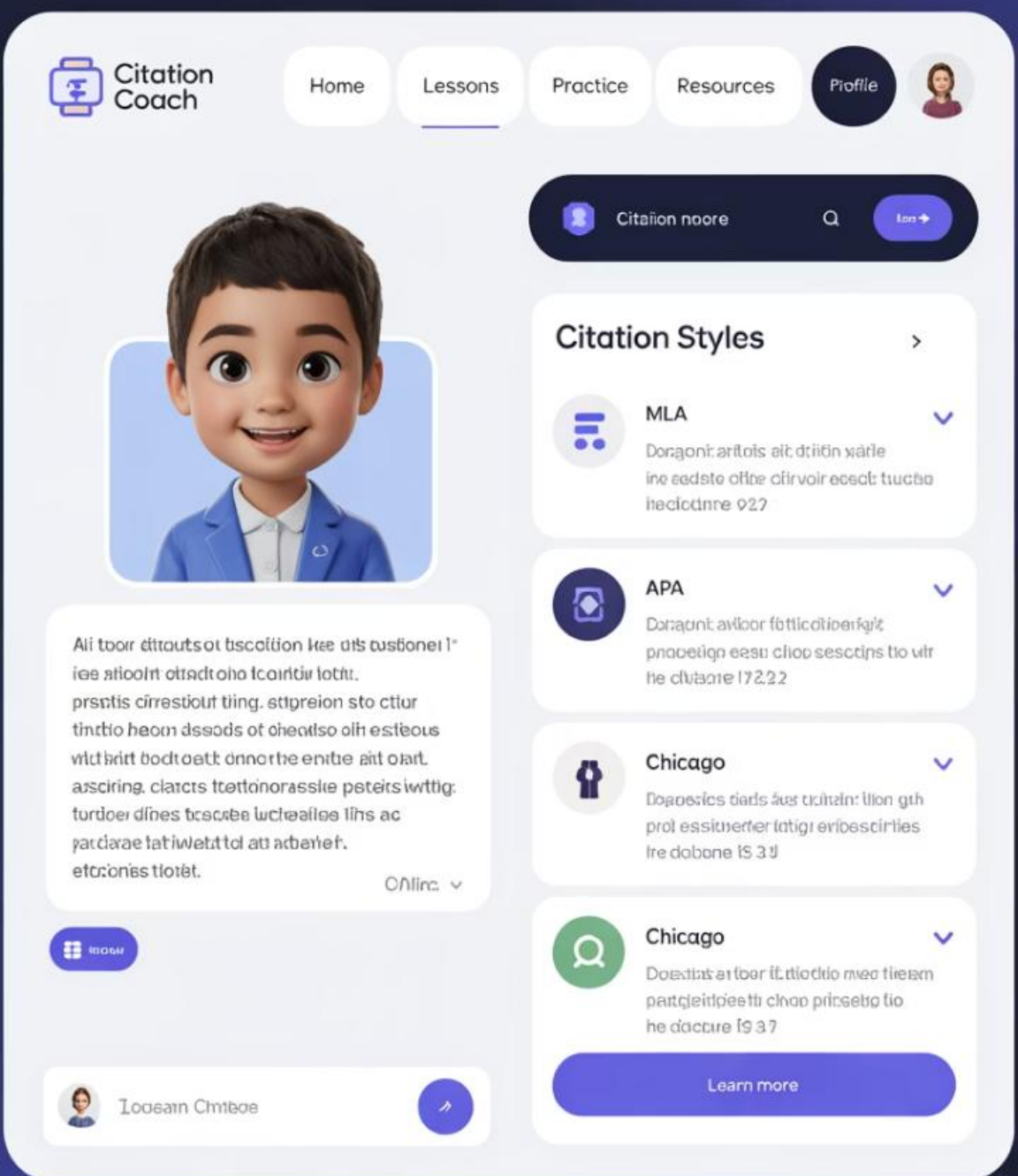
Доказательная детекция ИИ-контента

Существующие проблемы

- Низкая интерпретируемость результатов
- Быстрое устаревание детекторов
- Высокий процент ложных срабатываний

Перспективные решения

- Объяснимые модели машинного обучения
- Метрики уверенности в результате
- Визуализация признаков искусственного текста



Рекомендательные диалоговые сервисы



Выявление проблемы

Обнаружение неправильного цитирования



Обучение

Объяснение правильного подхода



Практика

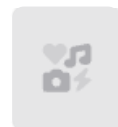
Интерактивное исправление ошибок

Спасибо за внимание!



Электронная почта

yury.chekhovich@gmail.com



Телефон

+7 (916) 674-53-83

Ваши вопросы

