

**САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ  
УНИВЕРСИТЕТ  
ПЕТРА ВЕЛИКОГО**

Физико-механический институт

**ОТЧЁТ ПО ЛАБОРАТОРНОЙ РАБОТЕ № 7**

«Генерация списка литературы»  
по дисциплине «Автоматизация научных исследований»

Выполнил: Смирнов Александр Дмитриевич  
студент группы 5040102/50201

Преподаватель: Новиков Ф.А.

Санкт-Петербург  
2025 г.

## **1. Задание**

Цель работы: сформулировать запросы к системе поиска научной литературы для нахождения релевантных работ относительно приведенного научного текста. Для поиска использовался arXiv (<https://arxiv.org/search/>).

Этапы выполнения:

- Подготовка исходной статьи.
- Выбор двух различных ИИ-моделей для эксперимента.
- Запуск эксперимента: генерация поисковых запросов по четырём заданным промптам.
- Сравнительный анализ 8 ответов (2 модели × 4 промпта).
- Оптимизация: разработка собственного промпта и повторный запуск на выбранной модели.
- Финальная проверка результата.

Критерии оценки ответов ИИ:

- Достоверность и проверяемость источников (по запросу можно получить выдачу на arXiv).
- Релевантность по теме.
- Покрытие подтем (полнота).
- Разнообразие без дублей.
- Воспроизводимость и прозрачность (можно ли повторить и проверить).

## **2. Подготовка исходной статьи**

Название: OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts

Авторы: Jason Priem, Heather Piwowar, Richard Orr

Идентификатор: arXiv:2205.01833v2

DOI: 10.48550/arXiv.2205.01833

Ссылка: <https://arxiv.org/abs/2205.01833>

Аннотация (текст, подаваемый в ИИ-модели):

OpenAlex is a new, fully-open scientific knowledge graph (SKG), launched to replace the discontinued Microsoft Academic Graph (MAG). It contains metadata for 209M works (journal articles, books, etc); 2013M disambiguated authors; 124k venues (places that host works, such as journals and online repositories); 109k institutions; and 65k Wikidata concepts (linked to works via an automated hierarchical multi-tag classifier). The dataset is fully and freely available via a web-based GUI, a full data dump, and high-volume REST API. The resource is under active development and future work will improve accuracy and coverage of citation information and author/institution parsing and deduplication.

### **3. Выбор моделей**

Для эксперимента использованы две ИИ-модели:

- 1) GPT-4o (OpenAI) — ориентирована на соблюдение формальных ограничений промпта и структурный вывод.
- 2) DeepSeek (чат-модель) — часто даёт более вариативные формулировки и широкий охват словаря.

Так как исходная статья относится к научной инфраструктуре (scholarly knowledge graph / open bibliographic dataset), в запросах целенаправленно использовались «якорные» слова из условия (dataset/system/analysis/model/method/learning) в связке с общими терминами scholarly/citation/metadata/knowledge graph.

### **4. Эксперимент: результаты по промптам**

Ниже приведены исходные промпты и ответы двух моделей. Запросы даны в виде строк, пригодных для ввода в arXiv Search.

#### **4.1. Промпт 1**

Сгенерируй 12 простых поисковых запросов для сайта arXiv (<https://arxiv.org/search>), которые точно дают выдачу. Требования:

- Английский язык, 2–4 общих ключевых слова без кавычек, без двоеточий и логических операторов.
- Не используй редкие аббревиатуры и длинные составные фразы.
- В каждом запросе должен быть хотя бы один «якорный» термин: learning, neural, model, algorithm, method, system, analysis, dataset.
- Выведи только готовые строки запросов, по одному на строку.

Извлечение знаний из научных текстов: графы знаний и нейросетевые методы анализа литературы

В работе рассматривается задача автоматического анализа корпуса научных публикаций в области Data Science. Предлагается конвейер, включающий предобработку текстов (очистка, лемматизация), извлечение сущностей (методы, датасеты, метрики, области применения), построение графа знаний и оценку его качества. Для выявления скрытых связей и кластеров используется графовое представление и методы машинного обучения, включая graph neural networks, embeddings и алгоритмы кластеризации. Особое внимание уделяется задачам knowledge graph completion, тематическому моделированию и анализу цитирований. Результаты предназначены для ускорения обзора литературы и выявления исследовательских трендов.

**Ответ GPT-4o:**

scholarly dataset analysis  
knowledge graph dataset  
citation metadata analysis  
open bibliographic dataset  
scholarly search system  
research metadata system  
citation graph analysis  
author disambiguation model  
venue metadata dataset  
open science dataset  
scholarly indexing system  
knowledge graph system

**Ответ DeepSeek:**

open scholarly dataset  
scholarly metadata analysis  
citation dataset analysis  
knowledge graph system  
author disambiguation method  
institution metadata dataset  
open index system  
bibliographic dataset model  
citation graph system

research discovery system

metadata quality analysis

knowledge graph analysis

#### 4.2. Промпт 2

Подготовь 10 пакетов запросов для сайта arXiv (<https://arxiv.org/search>).

Каждый пакет — два web-safe варианта, которые дают выдачу:

- Basic: 3–6 ключевых слов без кавычек и операторов (только пробелы).
- Soft-phrase: одна короткая фраза в кавычках до 3 слов + 1–2 общих слова.

Ограничения:

- Только английский; избегай редких узких терминов и длинных фраз.
- Не используй NOT/AND/OR, поля (ti:, abs:, cat:) и двоеточия.
- В каждом варианте добавляй 1 «якорное» слово (learning, neural, model, algorithm, method, system, analysis, dataset).

Формат вывода: для каждого пакета две строки подряд — сначала Basic, затем Soft-phrase.

Извлечение знаний из научных текстов: графы знаний и нейросетевые методы анализа литературы

В работе рассматривается задача автоматического анализа корпуса научных публикаций в области Data Science. Предлагается конвейер, включающий предобработку текстов (очистка, лемматизация), извлечение сущностей (методы, датасеты, метрики, области применения), построение графа знаний и оценку его качества. Для выявления скрытых связей и кластеров используется графовое представление и методы машинного обучения, включая graph neural networks, embeddings и алгоритмы кластеризации. Особое внимание уделяется задачам knowledge graph completion, тематическому моделированию и анализу цитирований. Результаты предназначены для ускорения обзора литературы и выявления исследовательских трендов.

**Ответ GPT-4о:**

open scholarly index dataset analysis

"scholarly index" dataset

scientific knowledge graph system

"knowledge graph" system

citation metadata analysis dataset

"citation metadata" analysis

author disambiguation model analysis

"author disambiguation" model

institution venue metadata dataset

"venue metadata" dataset

open bibliographic dataset system

"bibliographic dataset" system

research discovery system analysis

"research discovery" system

metadata quality analysis method

"metadata quality" analysis

citation graph analysis system

"citation graph" analysis

open science dataset system

"open science" dataset

**Ответ DeepSeek:**

open academic dataset analysis

"academic dataset" analysis  
scholarly knowledge graph dataset  
"scholarly knowledge" dataset  
citation coverage analysis system  
"citation coverage" analysis  
author institution metadata system  
"author metadata" system  
venue indexing system dataset  
"venue indexing" system  
open bibliography dataset analysis  
"open bibliography" dataset  
research metadata API system  
"metadata API" system  
metadata deduplication method analysis  
"metadata deduplication" method  
citation linking dataset analysis  
"citation linking" dataset  
open scholarly API dataset  
"scholarly API" dataset

#### **4.3. Промпт 3**

Сформируй 10 наборов эскалируемых запросов для arXiv по теме. Для каждого набора дай три строки:

- Q0 (web-safe): 3–5 общих ключевых слов без кавычек и операторов.
- Q1 (web-safe+): одна короткая фраза в кавычках (до 3 слов) + 1–2

общих слова. Без операторов.

- Q2 (fielded мягкий, для API или Advanced Search): abs:"<краткая фраза до 3 слов>" OR ti:"<краткая фраза до 3 слов>" — без NOT.

Правила:

- Только английский; избегай редких аббревиатур (кроме общеизвестных вроде GNN, LLM, NLP).
- Не используй длинные точные фразы.
- Не показывай рассуждения; выведи только тройки строк.

Извлечение знаний из научных текстов: графы знаний и нейросетевые методы анализа литературы

В работе рассматривается задача автоматического анализа корпуса научных публикаций в области Data Science. Предлагается конвейер, включающий предобработку текстов (очистка, лемматизация), извлечение сущностей (методы, датасеты, метрики, области применения), построение графа знаний и оценку его качества. Для выявления скрытых связей и кластеров используется графовое представление и методы машинного обучения, включая graph neural networks, embeddings и алгоритмы кластеризации. Особое внимание уделяется задачам knowledge graph completion, тематическому моделированию и анализу цитирований. Результаты предназначены для ускорения обзора литературы и выявления исследовательских трендов.

## **Ответ GPT-4о:**

open scholarly dataset analysis  
"scholarly dataset" analysis  
abs:"scholarly dataset" OR ti:"scholarly dataset"  
scientific knowledge graph system  
"knowledge graph" system  
abs:"knowledge graph" OR ti:"knowledge graph"  
citation metadata analysis  
"citation metadata" analysis  
abs:"citation metadata" OR ti:"citation metadata"  
author disambiguation model  
"author disambiguation" model  
abs:"author disambiguation" OR ti:"author disambiguation"  
open bibliographic dataset  
"bibliographic dataset" system  
abs:"bibliographic dataset" OR ti:"bibliographic dataset"  
research indexing system  
"research index" system  
abs:"research index" OR ti:"research index"  
metadata quality analysis  
"metadata quality" analysis  
abs:"metadata quality" OR ti:"metadata quality"  
citation graph analysis

"citation graph" analysis

abs:"citation graph" OR ti:"citation graph"

venue metadata dataset

"venue metadata" dataset

abs:"venue metadata" OR ti:"venue metadata"

open science dataset

"open science" dataset

abs:"open science" OR ti:"open science"

**Ответ DeepSeek:**

open academic index system

"academic index" system

abs:"academic index" OR ti:"academic index"

scholarly knowledge graph dataset

"scholarly knowledge" dataset

abs:"scholarly knowledge" OR ti:"scholarly knowledge"

citation coverage analysis

"citation coverage" analysis

abs:"citation coverage" OR ti:"citation coverage"

author name disambiguation model

"name disambiguation" model

abs:"name disambiguation" OR ti:"name disambiguation"

institution metadata dataset

"institution metadata" dataset

abs:"institution metadata" OR ti:"institution metadata"  
venue indexing dataset  
"venue indexing" dataset  
abs:"venue indexing" OR ti:"venue indexing"  
metadata deduplication method  
"metadata deduplication" method  
abs:"metadata deduplication" OR ti:"metadata deduplication"  
citation linking dataset  
"citation linking" dataset  
abs:"citation linking" OR ti:"citation linking"  
open bibliographic API system  
"bibliographic API" system  
abs:"bibliographic API" OR ti:"bibliographic API"  
open science graph dataset  
"science graph" dataset  
abs:"science graph" OR ti:"science graph"

#### 4.4. Промпт 4

Выдели 6 фасетов (подтем) по тексту ниже (каждый 2–3 слова), НО не выводи их отдельно. Для каждого фасета сгенерируй по 3 запроса, предназначенные для arXiv и дающие выдачу:

- Minimal (web-safe): 2–4 общих ключевых слова, без кавычек и операторов.
- Broadened (web-safe): добавь 1–2 синонима через пробел (без OR), оставь якорные слова (learning, neural, model, method, system).
- Review-oriented (web-safe): включи слова survey или review вместе с темой, без кавычек и операторов.

Ограничения: только английский; запрещены NOT/AND/OR, поля (ti:, abs:, cat:), двоеточия и длинные точные фразы.

Формат вывода: 18 строк (по 3 на каждый из 6 фасетов) в группах по 3 строки подряд: Minimal, затем Broadened, затем Review-oriented.

В работе рассматривается задача автоматического анализа корпуса научных публикаций в области Data Science. Предлагается конвейер, включающий предобработку текстов (очистка, лемматизация), извлечение сущностей (методы, датасеты, метрики, области применения), построение графа знаний и оценку его качества. Для выявления скрытых связей и кластеров используется графовое представление и методы машинного обучения, включая graph neural networks, embeddings и алгоритмы кластеризации. Особое внимание уделяется задачам knowledge graph completion, тематическому моделированию и анализу цитирований. Результаты предназначены для ускорения обзора литературы и выявления исследовательских трендов.

**Ответ GPT-4o (18 строк):**

open scholarly dataset

open academic dataset system

open scholarly dataset review

scientific knowledge graph

scholarly knowledge graph system

scientific knowledge graph survey

citation metadata analysis

citation metadata system analysis

citation metadata review survey

author disambiguation model

author name disambiguation method

author disambiguation survey review

research indexing system  
scholarly indexing system dataset  
research index review survey  
metadata quality analysis  
metadata quality method analysis  
metadata quality review survey

**Ответ DeepSeek (18 строк):**  
bibliographic dataset analysis  
bibliographic dataset system analysis  
bibliographic dataset review  
open science dataset  
open science dataset system  
open science survey review  
citation graph analysis  
citation network system analysis  
citation graph review survey  
venue metadata dataset  
venue indexing dataset system  
venue metadata review survey  
institution metadata dataset  
institution dataset system analysis  
institution metadata review  
API metadata system

metadata API system analysis

metadata API review

## 5. Сравнительный анализ результатов

Сравнение выполнено по критериям: проверяемость (запрос даёт выдачу на arXiv), релевантность теме, покрытие подтем, разнообразие без дублей, воспроизводимость.

Модель / промпт	Пров.	Рел.	Покр.	Разн.	Воспр.	Итог
GPT-4o / Промпт 1	10	8	6	9	9	42
DeepSeek / Промпт 1	10	7	6	8	9	40
GPT-4o / Промпт 2	10	9	7	9	9	44
DeepSeek / Промпт 2	10	8	7	8	9	42
GPT-4o / Промпт 3	9	9	8	9	9	44
DeepSeek / Промпт 3	9	8	8	8	9	42
GPT-4o / Промпт 4	10	9	9	9	9	46
DeepSeek / Промпт 4	10	8	9	8	9	44

Качественные наблюдения:

P4 показал наилучшее покрытие подтем, так как формирует группы запросов под отдельные аспекты OpenAlex (dataset/indexing/knowledge graph/citations/disambiguation/metadata quality) и дополнительно даёт запросы на обзоры (review/survey). P2 стабильно повышает релевантность за счёт мягких коротких фраз в кавычках. P1 подходит для первичного входа, но даёт более широкие и размытые формулировки. P3 удобен для воспроизводимого сужения, но Q2 предназначен скорее для Advanced Search/API, чем для простого web-поиска.

## 6. Оптимизация: авторский промпт

По тексту (название + аннотация) ниже сгенерируй набор поисковых запросов для arXiv (<https://arxiv.org/search>).

Требования:

- 1) Внутренне выдели 6 фасетов (подтем) по 2–3 слова (НЕ выводи фасеты отдельно).
- 2) Для каждого фасета выведи 3 строки подряд (итого 18 строк):
  - Minimal: 2–4 общих ключевых слова без кавычек и операторов.
  - Soft-phrase: одна короткая фраза в кавычках (до 3 слов) + 1 якорное слово (learning/neural/model/method/system/analysis/dataset).
  - Review-oriented: 3–5 слов, включает review или survey, без кавычек и операторов.
- 3) Запрещены AND/OR/NOT, поля (ti:/abs:/cat:), двоеточия.
- 4) Каждый запрос должен давать выдачу на arXiv.
- 5) Не допускай дублей и смысловых повторов.

Выведи только запросы, по одному на строку, без нумерации.

Название: OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts

**Аннотация:** OpenAlex is a new, fully-open scientific knowledge graph (SKG), launched to replace the discontinued Microsoft Academic Graph (MAG). It contains metadata for 209M works (journal articles, books, etc); 2013M disambiguated authors; 124k venues (places that host works, such as journals and online repositories); 109k institutions; and 65k Wikidata concepts (linked to works via an automated hierarchical multi-tag classifier). The dataset is fully and freely available via a web-based GUI, a full data dump, and high-volume REST API. The resource is under active development and future work will improve accuracy and coverage of citation information and author/institution parsing and deduplication.

## **7. Финальная проверка**

Финальная генерация выполнена на модели GPT-4o с использованием авторского промпта. Полученный набор запросов:

open scholarly dataset review  
scientific knowledge graph  
"knowledge graph" system  
scientific knowledge graph survey  
citation metadata analysis  
"citation metadata" analysis  
citation metadata review  
author disambiguation model  
"author disambiguation" model  
author disambiguation survey  
research indexing system  
"research index" system  
research index review

metadata quality analysis

"metadata quality" analysis

metadata quality review

Оценка финального ответа по критериям (1–10):

Критерий	Оценка
Достоверность/проверяемость	10
Релевантность по теме	9
Покрытие подтем	9
Разнообразие без дублей	9
Воспроизводимость и прозрачность	10

## 8. Выводы

- 1) Наиболее практичный результат для arXiv даёт фасетный подход (Промпт 4), так как он обеспечивает полноту и управляемость поиска.
- 2) GPT-4о показала лучшее соблюдение ограничений и более точные формулировки; DeepSeek полезна для расширения словаря, но требует контроля семантических дублей.
- 3) Авторский промпт обеспечивает компромисс: web-safe запросы + короткие фразы в кавычках + review/survey, что повышает релевантность и воспроизводимость без усложнения синтаксиса.