

**САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
ПЕТРА ВЕЛИКОГО**

Физико-механический институт

ОТЧЁТ ПО ЛАБОРАТОРНОЙ РАБОТЕ № 6

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

Выполнил: Смирнов Александр Дмитриевич
студент группы 5040102/50201
Преподаватель: Новиков Ф.А.

Санкт-Петербург
2025 г.

Тема исследования: Автоматическое построение графа знаний по корпусу научных публикаций в области Data Science для выявления трендов и ключевых работ.

Краткое описание: Для проведения эксперимента была задана тематическая область, связанная с анализом научных публикаций и построением графов знаний в области Data Science, на основе которой выполнялась генерация библиографических источников.

1. Цель и постановка эксперимента

Цель: оценить эффективность двух нейросетевых моделей при формировании списка литературы по заданной теме. План эксперимента: 2 модели \times 3 промпта (P1-P3) \times 3 источника = 18 источников.

Модели: (1) GPT-5.2 Thinking; (2) GPT-4o.

Для каждого промпта отбирались первые три уникальные ссылки, после чего выполнялась верификация и выставление баллов по критериям

2. Используемые промпты

P1: Сгенерируй список научных источников, которые могут быть полезны для исследования данной темы. Укажи URL. Учитывай индекс цитируемости. Автоматическое построение графа знаний по корпусу научных публикаций в области Data Science для выявления трендов и ключевых работ. Работа посвящена подбору релевантных научных источников для исследования методов построения графов знаний по научным публикациям. Рассматриваются подходы NLP (извлечение сущностей и отношений из научных текстов), анализ сетей цитирования и представления графов (эмбединги), а также использование открытых корпусов научной литературы (S2ORC, OpenAlex).

P2: Используя предоставленное название и краткое описание статьи, составь список релевантных научных источников. Укажи только реально существующие публикации: статьи, книги, мета-обзоры или авторитетные конференционные/журнальные работы. Отбирай источники по тематической близости, теоретической значимости и актуальности. Укажи валидные URL. Учитывай индекс цитируемости. Автоматическое построение графа знаний по корпусу научных публикаций в области Data Science для выявления трендов и ключевых работ. Работа посвящена подбору релевантных научных источников для исследования методов построения графов знаний по научным публикациям. Рассматриваются подходы NLP (извлечение сущностей и отношений из научных текстов), анализ сетей цитирования и представления графов (эмбединги), а также использование открытых корпусов научной литературы (S2ORC, OpenAlex).

P3: Проанализируй предоставленные название и краткое описание научной темы. На основе заданной тематики сформируй точный и проверяемый список литературы, состоящий исключительно из реально существующих научных публикаций. Не изобретай источники. Обязательно проверяй все ссылки, которые выдаёшь. Включай только работы из рецензируемых журналов, научных книг, конференций или авторитетных репозиториях. Отбирай литературу по следующим критериям: строгая релевантность теме, фундаментальность и/или современность исследования, практическая ценность, разнообразие типов источников, индекс цитируемости. Выводи сведения об авторах, годе публикации, точном названии, журнале или конференции (либо издатель), а также DOI при его наличии. Ко всем работам указывай валидные ссылки, по которым доступен полный текст.

Заданная тема исследования:

Методы автоматического формирования графов знаний на основе научных публикаций в области Data Science.

В рамках лабораторной работы указанная тема используется в качестве входных данных для эксперимента по автоматической генерации библиографического списка. Нейросетевые модели подбирают релевантные научные источники, связанные с анализом научных публикаций, методами обработки естественного языка, построением графов знаний и анализом сетей цитирования.

3. Полученные источники (по 3 на каждый промпт и модель)

GPT-5.2 Thinking

P1

- Hogan A., Blomqvist E., Cochez M. et al. Knowledge Graphs // ACM Computing Surveys. 2021. Vol. 54, No. 4, Art. 71. DOI:10.1145/3447772. URL: <https://arxiv.org/abs/2003.02320>
- Lo K., Wang L. L., Neumann M. et al. S2ORC: The Semantic Scholar Open Research Corpus // ACL 2020. URL: <https://aclanthology.org/2020.acl-main.447/>
- Priem J., Piwowar H., Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. 2022. arXiv:2205.01833. URL: <https://arxiv.org/abs/2205.01833>.

P2

- Grover A., Leskovec J. node2vec: Scalable Feature Learning for Networks. 2016. arXiv:1607.00653. URL: <https://arxiv.org/abs/1607.00653>
- Hamilton W. L., Ying R., Leskovec J. Inductive Representation Learning on Large Graphs (GraphSAGE). 2017. arXiv:1706.02216. URL: <https://arxiv.org/abs/1706.02216>
- Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. 2013. arXiv:1301.3781. URL: <https://arxiv.org/abs/1301.3781>.

P3

- Hogan A., Blomqvist E., Cochez M. et al. Knowledge Graphs // ACM Computing Surveys. 2021. Vol. 54, No. 4, Art. 71. DOI:10.1145/3447772. URL: <https://dl.acm.org/doi/10.1145/3447772>.
- Lo K., Wang L. L., Neumann M. et al. S2ORC: The Semantic Scholar Open Research Corpus. 2019. arXiv:1911.02782. URL: <https://arxiv.org/abs/1911.02782>.
- Priem J., Piwowar H., Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. 2022. arXiv:2205.01833. URL: <https://arxiv.org/pdf/2205.01833>.

GPT-4o

P1

- Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. Vol. 3. P. 993–1022. URL: <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Beltagy I., Lo K., Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. 2019. arXiv:1903.10676. URL: <https://arxiv.org/abs/1903.10676>.
- W3C. RDF 1.1 Concepts and Abstract Syntax. 2014. URL: <https://www.w3.org/TR/rdf11-concepts/>.

P2

- Luan Y., He L., Ostendorf M., Hajishirzi H. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction // EMNLP 2018. URL: <https://aclanthology.org/D18-1360/>.
- Microsoft Research. Microsoft Academic Graph (MAG) — проект и описание набора данных. 2015. URL: <https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/>.
- W3C. SPARQL 1.1 Query Language. 2013. URL: <https://www.w3.org/TR/sparql11-query/>.

P3

- Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Journal of Machine Learning Research. 2003. Vol. 3. P. 993–1022. URL: <https://jmlr.csail.mit.edu/papers/v3/blei03a.html>.
- Beltagy I., Lo K., Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. 2019. arXiv:1903.10676. URL: <https://arxiv.org/pdf/1903.10676>.
- Sinha A., Shen Z., Song Y. et al. An Overview of Microsoft Academic Service (MAS) and Applications // WWW '15 Companion. 2015. URL: <https://www.microsoft.com/en-us/research/project/open-academic-graph/>.

4. Сводная таблица оценок (18 источников)

Критерии: Существование, Оформление(ГОСТ), Полезность, Новизна, Индекс Цитируемости (Q1-Q4), Разнообразие типа.

№	Модель	Промпт	Тип	Год	URL	Сущ.	ГОСТ	Полез.	Нов.	Цит.	Разн.	Итого
1	GPT-5.2 Thinking	P1	Статья (журнал)	2021	ТЫК	10	7	10	10	4	7	48
2	GPT-5.2 Thinking	P1	Статья (конф.)	2020	ТЫК	10	7	9	10	7	7	50
3	GPT-5.2 Thinking	P1	Препринт (arXiv)	2022	ТЫК	10	7	9	10	4	9	49
4	GPT-5.2 Thinking	P2	Препринт (arXiv)	2016	ТЫК	10	8	8	10	4	9	49
5	GPT-5.2 Thinking	P2	Препринт (arXiv)	2017	ТЫК	10	8	8	10	4	9	49
6	GPT-5.2 Thinking	P2	Статья (журнал)	2013	ТЫК	10	8	7	10	4	7	46
7	GPT-5.2 Thinking	P3	Статья (журнал)	2021	ТЫК	10	7	10	10	4	7	48
8	GPT-5.2 Thinking	P3	Статья (конф.)	2020	ТЫК	10	7	9	10	7	7	50
9	GPT-5.2 Thinking	P3	Препринт (arXiv)	2022	ТЫК	10	7	9	10	4	9	49
10	GPT-4o	P1	Статья (журнал)	2003	ТЫК	10	8	7	8	10	7	50
11	GPT-4o	P1	Препринт (arXiv)	2019	ТЫК	10	7	8	10	4	9	48
12	GPT-4o	P1	Спецификация (стандарт)	2014	ТЫК	10	7	6	10	4	9	46
13	GPT-4o	P2	Статья (конф.)	2018	ТЫК	10	8	9	10	7	7	51
14	GPT-4o	P2	Веб-ресурс (проект датасет)	2015	ТЫК	10	8	8	10	5	8	49
15	GPT-4o	P2	Спецификация (стандарт)	2013	ТЫК	10	8	7	10	4	9	48
16	GPT-4o	P3	Статья (журнал)	2003	ТЫК	10	8	7	8	10	7	50
17	GPT-4o	P3	Препринт (arXiv)	2019	ТЫК	10	7	8	10	4	9	48
18	GPT-4o	P3	Статья (конф.)	2015	ТЫК	9	10	8	10	5	7	49

5. Итоговые результаты и сравнительный анализ

Средние значения по моделям:

Средние по модели	Сущ.	ГОСТ	Полез.	Нов.	Цит.	Разн.	Итого
GPT-4o	9,89	7,89	7,56	9,56	6,44	8,00	49,34
GPT-5.2 Thinking	10,00	7,33	8,78	10	4,67	7,89	48,67

Средние значения по промптам:

Средние по промпту	Сущ.	ГОСТ	Полез.	Нов.	Цит.	Разн.	Итого
P1	10,00	7,17	8,17	9,67	5,50	8	48,50
P2	10,00	8,00	7,83	10,00	4,67	8,17	48,67
P3	9,84	7,67	8,5	9,67	5,67	7,67	49,02

Анализ средних значений показывает, что различия между моделями и уровнями сложности промптов носят умеренный характер и не сводятся к однозначному доминированию одной стратегии генерации.

При сравнении моделей видно, что GPT-4o демонстрирует немного более высокий итоговый средний балл (49,34 против 48,67). Это преимущество формируется прежде всего за счёт более высоких значений по критериям «Оформление по ГОСТ», «Индекс цитируемости» и «Разнообразие типов источников». GPT-4o чаще предлагает журнальные публикации из высокорейтинговых изданий (Q1–Q2), что положительно сказывается на показателе цитируемости. В то же время GPT-5.2 Thinking показывает более высокие значения по критериям «Полезность» и «Новизна», предлагая более современные и тематически близкие источники, однако это сопровождается снижением среднего балла по индексу цитируемости за счёт включения препринтов и инфраструктурных материалов.

Сравнение промптов показывает, что рост сложности формулировки не приводит к линейному улучшению итогового качества. Промпт P1 уже обеспечивает высокий уровень по большинству критериев и демонстрирует стабильные значения, особенно по критериям «Существование» и «Новизна». Промпт P2 улучшает показатели оформления, однако сопровождается снижением среднего балла по критерию цитируемости. Наиболее высокий итоговый балл получен у промпта P3 (49,02), что объясняется более сбалансированным распределением оценок между полезностью, оформлением и разнообразием типов источников, несмотря на небольшое снижение по критерию существования из-за большего числа ссылок на платные издательские платформы.

6. Вывод

В целом эксперимент показывает, что детализация промпта повышает структурированность и академическую аккуратность результатов, однако эффект не является радикальным. Качество генерируемых списков определяется не только строгостью запроса, но и особенностями модели: одни модели ориентированы на авторитетность и цитируемость источников, другие — на актуальность и практическую полезность. Это подтверждает целесообразность использования нейросетевых моделей как инструмента предварительного подбора литературы с последующей экспертной проверкой и доработкой.