МФТИ

Разработка интеллектуальной системы анализа патентов химической отрасли для представления данных в структурированном виде

Выполнил: Кайда Анатолий Сергеевич

Научный руководитель: Глинский Андрей Владимирович

ЦЕЛЬ ИССЛЕДОВАНИЯ

ЦЕЛЬ:

Определить методы извлечения информации о процедурах синтеза гетерогенных катализаторов с попощью современных больших языковых моделей (LLM)

ОБЪЕКТ:

Современные LLM в задаче извлечения качественной стркутурированной инорфмации из патентной документации

ПРЕДМЕТ:

Изучение совеременных LLM в задаче извлечения информации из патентов в домене «катализаторы синтеза полиолефинов»

ПРОБЛЕМА:

▶ Количество ежегодно публикуемых статей и патентов в химии растет экспоненциально

▶ Процесс работы с патентной и литературной информацией по-прежнему остается в значительной степени ручным

▶ Навигации в большом объеме литературы остается чрезвычайно сложной. Поэтому важные научные открытиостаются незамеченными в течение длительного времени

▶ Отсутствуют открытые сервисы и решения позволяющие извлекать сложную контекстную информацию из патентных документов

▶ Большинство информации хранится в неструктрированном виде.

АКТУАЛЬНОСТЬ:

▶ Большинство промышленных процессов в химической промышленности основаны на использовании катализаторов (более 85% всех известных процессов)

▶ Полиолефины являются саммым крупнотонажным искусственным полимером на сегодняшний день

▶ С релизом GPT-3.5 наблюдается рост публикаций описывающих применение LLM для решения задач в области естественных наук.

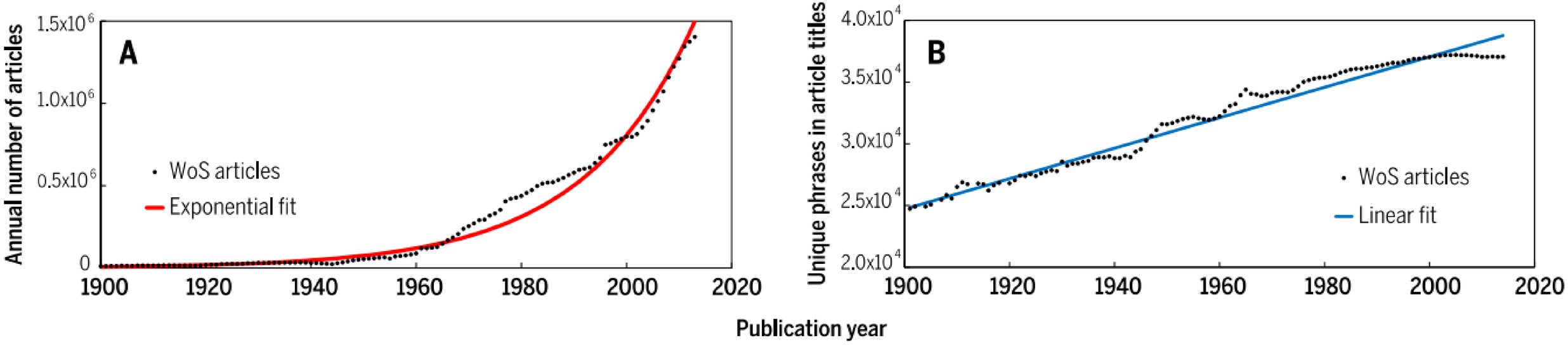
НОВИЗНА:

▶ Разработан подход для высокоточного извлечения экспериментальной информации о процессе синтеза из документов патентных документов в доменной области «катализатора синтеза полиолефинов»

ГИПОТЕЗА

Современные LLM применимы для создания высокоточных системы для извлечения сложной неструктурированной информации из научного текста в домене «катализаторы синтеза полиолефинов»

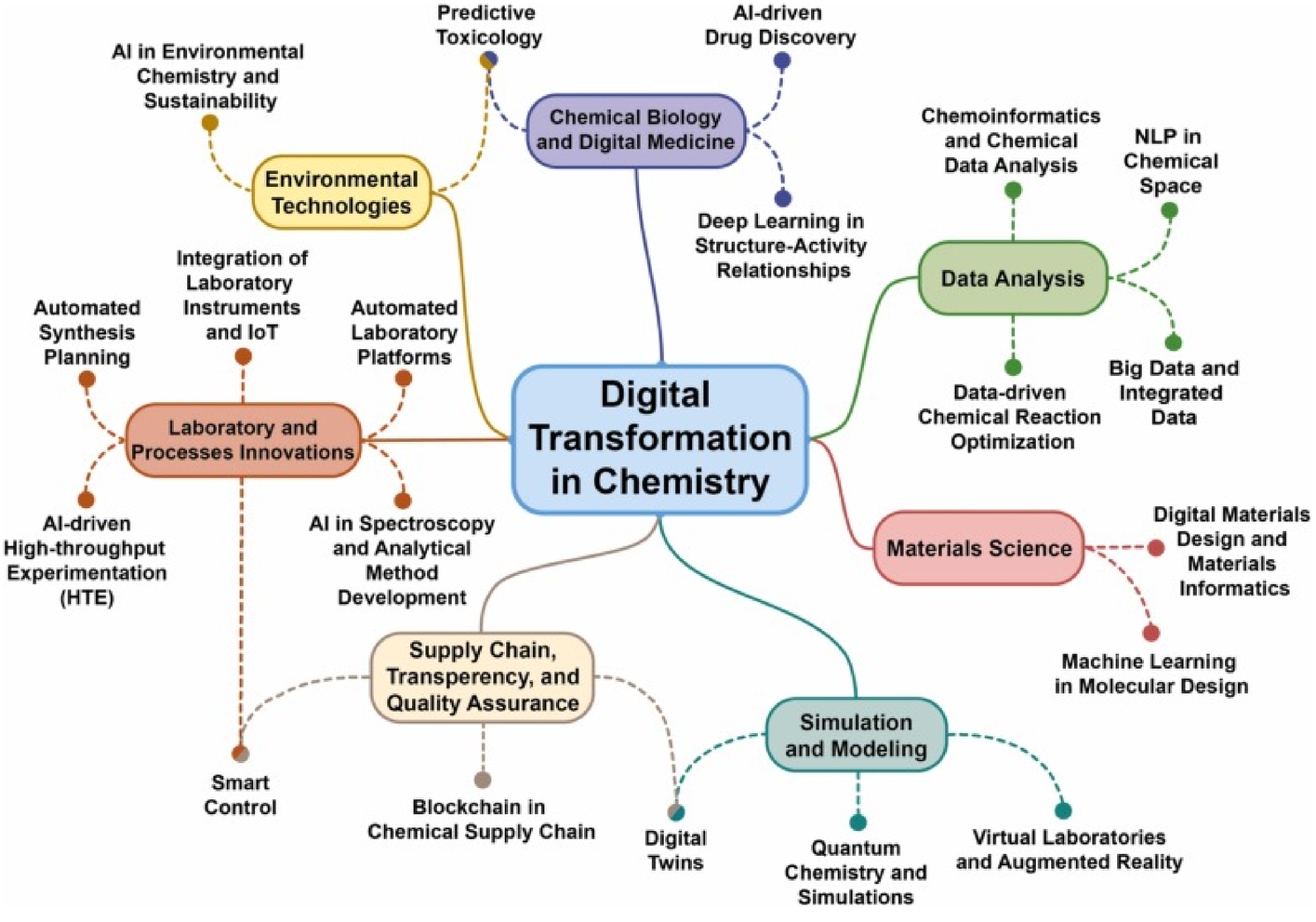
Литературный обзор



1. Science of science | Science [Electronic resource]. URL: https://www.science.org/doi/10.1126/science.aao0185 (accessed: 01.11.2024).

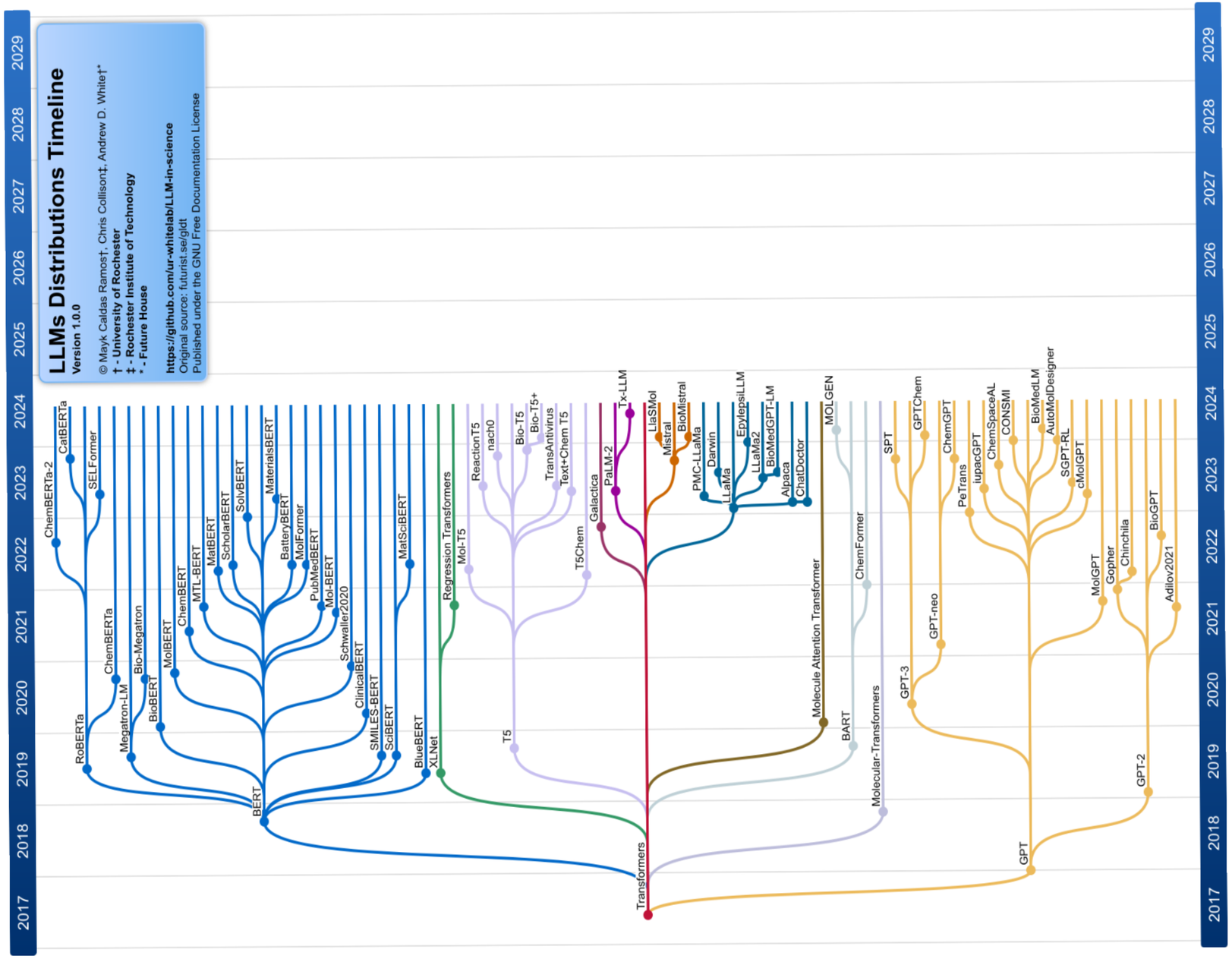
Литературный обзор

2. Ananikov V. Top 20 Influential AI-Based Technologies in Chemistry. Chemistry, 2024.



В настоящее время используется только 10% доступной в литературе химической информации!

Литературный обзор

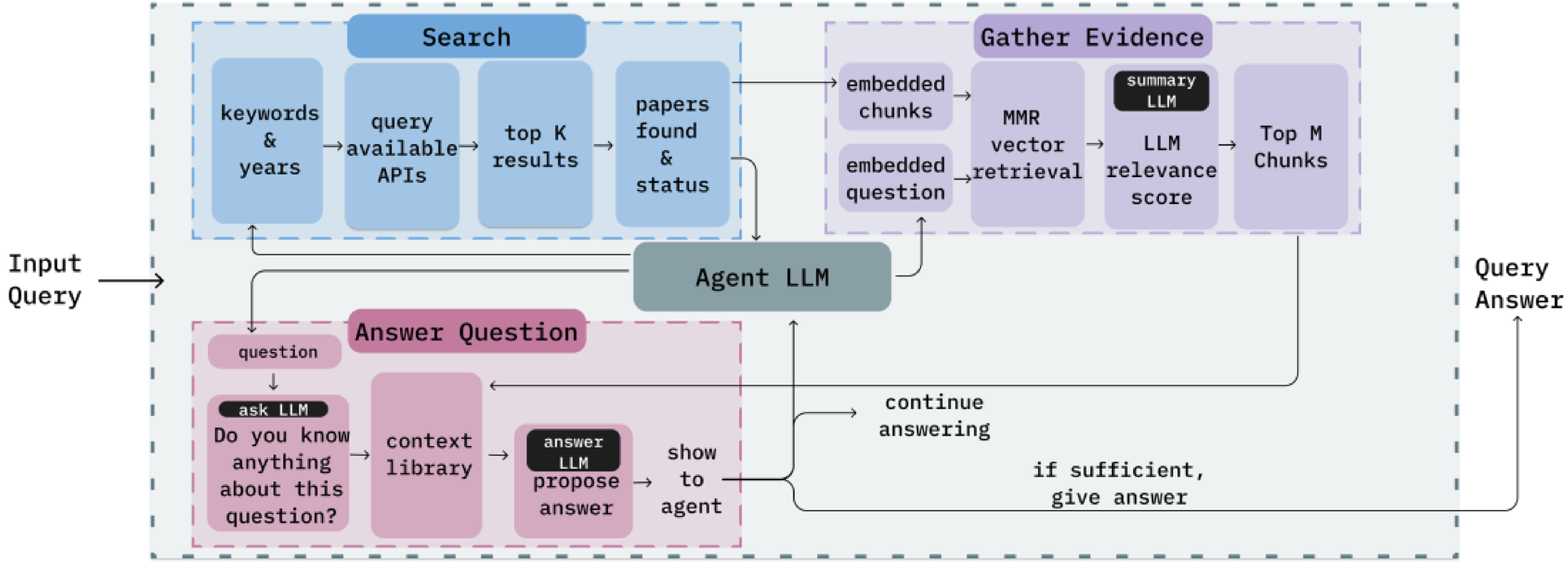


3. Ramos M.C., Collison C.J., White A.D. A Review of Large Language Models and Autonomous Agents in Chemistry: arXiv:2407.01603. arXiv, 2024.

С появлением «Трансформеров» наблюдается «бум» в развитии в создании специализированных моделей и сервисов для решения специфических задач химии

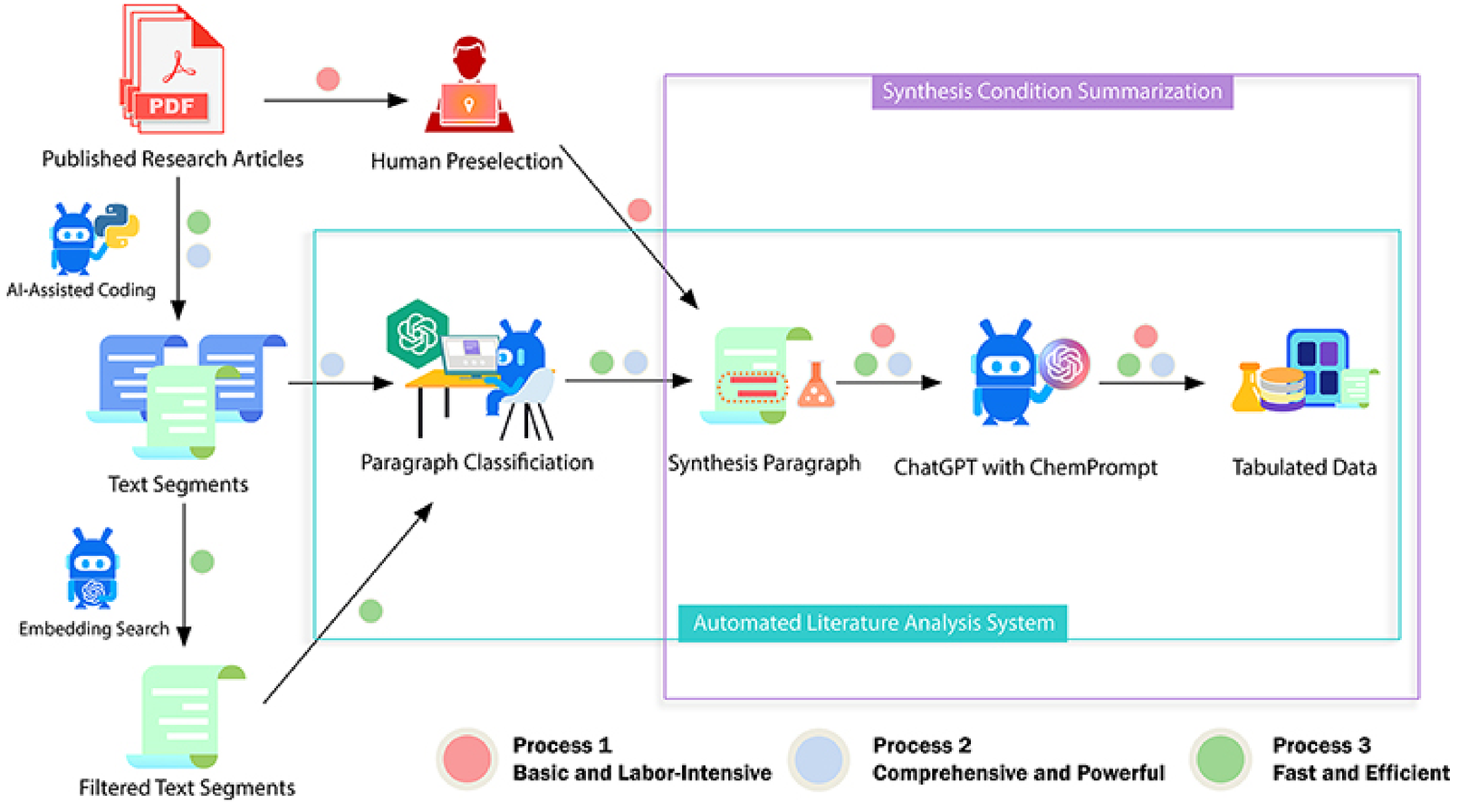
Литературный обзор

4. Lála J. et al. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research: arXiv:2312.07559. arXiv, 2023.



PaperQA — это агент, который преобразует вопрос в ответ с указанием источников. Агент использует три инструмента: поиск, сбор данных и ответ на вопрос. Инструменты позволяют ему находить и анализировать соответствующие полнотекстовые исследовательские работы, определять конкретные разделы в работе, которые помогают ответить на вопрос, суммировать эти разделы с контекстом вопроса (называемые доказательствами), а затем генерировать ответ на основе доказательств.

Литературный обзор



5. Zheng Z. et al. ChatGPT Chemistry Assistant for Text Mining and Prediction of MOF Synthesis.

В течении 2024 начали активно появляться публикации о применении LLM для извлечения экспериментальной информации из публикаций описывающих различные химические процедуры

ЗАДАЧИ ИССЛЕДОВАНИЯ

▶ Выбрать домен для проведения исследования и провести релевантный патентный поиск и создать БД документов для дальнейшего извлечения информации

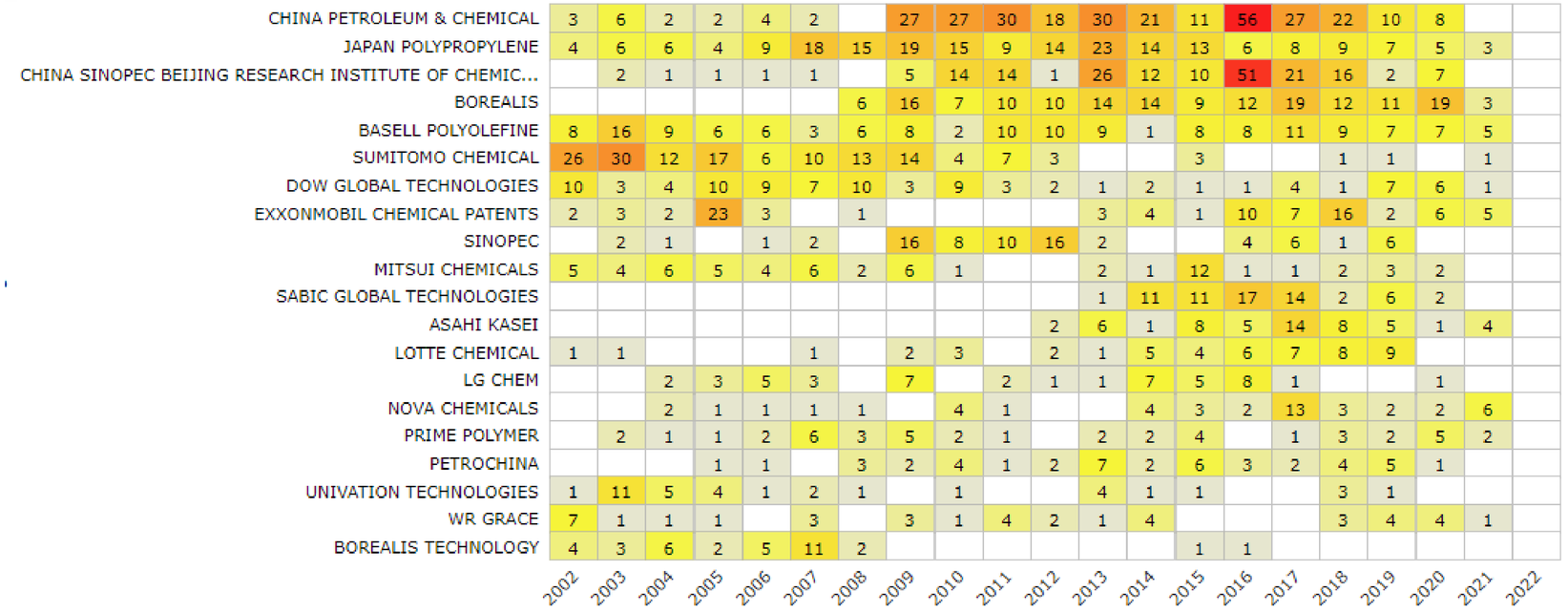
▶ Провести обзор современных фреймворков и технологий для работы с LLM

▶ Создать систему на основе LLM способную решать следующие задачи:

- Сегментация и фильтрация текста
- Классификация текста
- Выделение информации о синтетических процедурах
- Определение структуры текста, построение графа ссылок сегментов текста друг на друга
- Выделение необходимо информации
- Запись информации в БД

ИСТОЧНИКИ ДАННЫХ И ОСОБЕННОСТИ ДОМЕНА

1)Catalyst type:	**1)PE: ZN on SiO2**
1)KEY WORDS:	**1)T/A/C/D: (((((Ziegler-Natta catalyst+) or (silica?supported Ziegler-Natta catalyst+)) and (titanium +chloride)) and gas-phase) and (PE or polyethylene))**
1)RESTRICTION	**1)ALL**
1)number of documents before relevance is determined	**1)There were about 2491 documents (ORBIT)**
1)DATE	**1)Priority date from 01/01/2002**
1)2491	1)patented inventions
1)0,5	1)owned by top 10 players



Домен выбранный для исследования относится к широкому классу «гетерогенных катализаторов». Для данного домена можно выделить следущие особенности:

1) Практически вся передовая и коммерически ценная информация содержится в патентной литературе;

2) Практически отсутствуют способов описать катализаторов в терминах «классической» химии;

3) Высокая стоимость одной экспериментальной точки. Цена одного эксперимента может достигать нескольких тысяч долларов;

4) Практически отсутсвуют стандарты при написании патентов, структура экспериментальной части для каждого отдельного документа зачастую уникальна.

ПРИМЕР ИСТОЧНИКА ДАННЫХ

1) Информация о катализаторе представлена в виде «рецепта» и набора нескоторных свойств;

2) Как видно в примере, часть экспериментов ссылаются друг надруга в документе, при этом сам документ не является размеченным;

3) Часть важной информации описывающей свойства катализатора может находится в разных частях документа и таблицах;

4) Размерности физических величин зачастую отличаются в зависимости от года, региона и т. д.

Example 3

[0092] In a first step, 40.9 g of finely divided spray-dried silica gel ES 70X from Crossfield, which had been dried at 600° C., were suspended in ethylbenzene and admixed while stirring with 2.7 ml of diethylaluminum chloride (2 M in heptane). 57.3 ml of $(n\text{-butyl})_{1.5}(octyl)_{0.5}$ magnesium (0.875 M in n-heptane) were then added. 11.45 ml of tert-butyl chloride were added to the solid obtained-in this way and a solution of 1 ml of ethanol was then slowly added dropwise. 5.5 ml of titanium tetrachloride were added to this mixture, the resulting solid was filtered off, resuspended in pentane, and 5.18 ml of hexamethyldisilazane were then added. The pentane was distilled off and the catalyst system obtained in this way was dried under reduced pressure. This gave 70.3 g of the catalyst system according to the present invention.

Example 4 (Comparative Example)

[0093] The preparation of the catalyst was carried out using the same components in the same mass and molar ratios as in example 1, but without addition of diethylaluminum chloride (step A).

Example 5 (Comparative Example)

[0094] The preparation of the catalyst was carried out using the same components in the same mass and molar ratios as in example 1, but without addition of ethanol (step D).

Example 6 (Comparative Example)

[0095] The preparation of the catalyst was carried out using the same components in the same mass and molar ratios as in example 2, but without addition of ethanol (step D).

Example 7 (Comparative Example)

[0096] In a first step, 25.7 g of finely divided spray-dried silica gel ES 70X from Crossfield, which had been dried at 600° C., were suspended in ethylbenzene and admixed while stirring with 1.7 ml of diethylaluminum chloride (2 M in heptane). 36 ml of $(n\text{-butyl})_{1.5}(octyl)_{0.5}$ magnesium (0.875 M in n-heptane) were then added. 5.51 ml of chloroform were added to the solid obtained in this way and a solution of 0.83 ml of tetrahydrofuran was then slowly added dropwise. 3.4 ml of titanium tetrachloride were added to this mixture, the resulting solid was filtered off, resuspended in pentane, and 3.25 ml of hexamethyldisilazane were then added. The pentane was distilled off and the catalyst system obtained in this way was dried under reduced pressure. This gave 33.9 g of the catalyst system.

Examples 8 to 11

[0097] Polymerization

reports the productivity of the catalyst systems from examples 1 to 4 both for the examples 8 to 10 according to the present invention and for the comparative example 11.

TABLE 1

Polymerization results

Ex.	Catalyst from ex.	Weight of catalyst [mg]	Polymerization time [min]	Yield [g of PE]	Productivity [g of PE/ g of cat]
8	1	38	120	270	7105
9	2	99	60	450	4545
10	3	132	60	110	833
11	4 (C)	47	120	210	4468

Examples 12 and 13

[0099] Polymerization

[0100] The polymerizations were carried out under the same conditions as described in examples 8 to 11 using the catalysts from example 3 and comparative example 5. The catalyst from example 3 gave an ethylene copolymer having a bulk density of 416 g/l. The catalyst from comparative example 5 gave an ethylene copolymer having a bulk density of 195 g/l.

Examples 14 to 16

[0101] Polymerization

[0102] 200 mg of triisobutylaluminum were introduced into a 10 l autoclave which had been charged with 150 g of polyethylene and made inert by means of argon. The autoclave was then pressurized with 1 bar of H_2 and 10 bar of ethylene, the weight of catalyst indicated in table 2 was added and polymerization was carried out at an internal reactor temperature of 110° C. for one hour. The reaction was stopped by venting.

[0103] Table 2 below reports the productivity of the catalyst systems used and the bulk densities of the ethylene polymers obtained both for examples 14 and 15 according to the present invention and for the comparative example 16.

TABLE 2

Polymerization results

Ex.	Catalyst from ex.	Weight of catalyst [mg]	Bulk density [g/l]	Yield [g of PE]	Productivity [g of PE/ g of cat]
14	1	112	249	124	1107
15	2	82	358	104	1268
16	6 (C)	92	324	85	924

Современные фреймворки и технологии для работы с LLM

Фреймворк	Задача	Недостатки	Преимущества
LangChain	Создание и развертывание приложений с LLM, упрощение интеграции моделей	Ограниченная масштабируемость, сложности в отладке	Быстрое прототипирование, модульная архитектура, открытый исходный код
LangGraph	Управление сложными рабочими процессами с LLM, поддержка циклических графов	Сложность настройки, ограничения масштабируемости	Управление сложными рабочими процессами с LLM, поддержка циклических графов
CrewAI	Создание AI-агентов для автоматизации задач, поддержка различных LLM	Высокая сложность отладки и настройки системы, ограничен доступ для неквалифицированных пользователей, высокая стоимость некоторых тарифных планов	Открытый исходный код, высокая настраиваемость, поддержка множества LLM
OpenAI API	Интеграция AI в приложения с помощью предобученных моделей, улучшение пользовательского опыта	Высокая стоимость на передовые модели, зависимость от модели, риски утечки данных, ограничения на количество запросов в единицу времени	Простой API, масштабируемая инфраструктура, доступ к передовым моделям, возможность использования структурированного вывода (совместимость с Pydantic)

Для построения «ядра» системы анализа патентов и проведения предварительных экспериментов решено использовать API OpenAI ввиду доступности и простоты работы с передовыми моделями семейства GPT-4o.

Эксперименты по извлечению необходимых данных из “сырого” текста

1) Большинство файлов патентов хранятся в фомате pdf с текстовый слоем и не имеют разметки;

2) Классические подходы в построении RAG систем не позволяют гарантировать качественное и полноценное извлечение данных

3) Современные модели семейства gpt-4o имеют большое контекстное окно, что потенциально позволяет обрабатывать большинство документов целиком;

4) API OpenAI представляет удобный интерфейс настройки структурированного вывода

5) Для изучения возможностей в первичной оценки качества работы моделей было выбрано 10 тестовых документа различной структуры и

```python
class TableData(BaseModel):
    model_config = ConfigDict(extra='forbid')

    headers: List[str] = Field(
        default_factory=list,
        description="Column headers of the table"
    )
    rows: List[List[str]] = Field(
        default_factory=list,
        description="Table data rows"
    )
    caption: Optional[str] = Field(
        default=None,
        description="Table caption or description"
    )

class ExperimentInfo(BaseModel):
    model_config = ConfigDict(extra='forbid')

    id: str = Field(description="Unique experiment identifier (e.g., 'Example 1')")
    text_of_example: str = Field(description="Exact text quote from the patent")
    type: str = Field(
        description="Type of experiment",
        # Using Field with allowed values instead of pattern
        json_schema_extra={"enum": ["catalyst_synthesis", "polymerization", "table"]}
    )
    reference: List[str] = Field(
        default_factory=list,
        description="List of referenced experiment IDs"
    )
    table_data: Optional[TableData] = Field(
        default=None,
        description="Structured table data if type is 'table'"
    )

class PatentInfo(BaseModel):
    model_config = ConfigDict(extra='forbid')

    experiments: List[ExperimentInfo] = Field(
        default_factory=list,
        description="List of experiments found in the patent"
    )
```

Результаты экспериментов с моделью GPT-4o-mini

На тестовой выборке из 10 документов модель gpt-4o-mini не смогла извлечь достоверные данные не из одного документа:

- Например, документа US4849389, содержащего описание 8 экспериментов и 3 таблицы не удалось извлечь достоверно извлечь только 2 эксперимента.
- Степень извлечения данных не превышает 25%. Это не позволяет использовать данные подход для качественного анализа текстов

Модель точно классифицирует найденные экспериментальные данные

Структура данных в ответе модели соответствует тробованиям

```
gpt_4o_mini_results = extract_info_agent(text=text, model="gpt-4o-mini")
pretty_print_patent_info(gpt_4o_mini_results)
```

```
2025-03-29 23:04:41,523 - INFO - HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"

=== Patent Analysis Results ===

{
  "experiments": [
    {
      "id": "Example 1",
      "text_of_example": "20.0 grams of Davison grade 955 silica which had been heated to 600 C. for about 16 ho
      "type": "catalyst_synthesis",
      "reference": [],
      "table_data": null
    },
    {
      "id": "Example 2",
      "text_of_example": "Eight (8) additional catalyst precursor compositions were synthesized in the manner sul
      "type": "table",
      "reference": [
        "Example 1"
      ],
      "table_data": {
        "headers": [
          "Example",
          "Alcohol Type",
          "Transition Metal",
          "Mg (MMOLS/GRAM)",
...
  ]
}

=== End of Analysis ===
```

Результаты экспериментов с моделью GPT-4o

На тестовой выборке из 10 документов модели gpt-4o показала значительно более лучший результат:

- Количество извлеченных данных значительно увеличилось;
- Для документа US4849389, содержащего описан 8 экспериментов и 3 таблицы удалось извлечь достоверно извлечь 7 экспериментов и одну таблицу достоверно.
- Структура данных заполняется корректно, корректно отображаются ссылки
- Модель точно классифицирует найденные экспериментальные данные

Наблюдается наличие воспроизводимых артефактов: дубликаты, галлюцинации

Таблицы стабильно некачественно


```
  ],
  "table_data": null
},
{
  "id": "Example 6",
  "text_of_example": "EXAMPLES 2-7\n (Catalyst Synthesis) \n Eight (8) additional catalyst precursor compositions \n were synthesized in the
  "type": "catalyst_synthesis",
  "reference": [
    "Example 1"
  ],
  "table_data": null
},
{
  "id": "Example 6A",
  "text_of_example": "EXAMPLES 2-7\n (Catalyst Synthesis) \n Eight (8) additional catalyst precursor compositions \n were synthesized in the
  "type": "catalyst_synthesis",
  "reference": [
    "Example 1"
  ],
  "table_data": null
},
{
  "id": "Example 7",
  "text_of_example": "EXAMPLES 2-7\n (Catalyst Synthesis) \n Eight (8) additional catalyst precursor compositions \n were synthesized in the
  "type": "catalyst_synthesis",
  "reference": [
    "Example 1"
  ],
  "table_data": null
},
{
  "id": "Example 8",
  "text_of_example": "EXAMPLES 8-14\n (Polymerization Process) \n The catalyst precursors of Examples 1-7 were com bined with triethylalumin
  "type": "polymerization",
  "reference": [
    "Example 1"
```

Предварительная оценка качество работы моделей в задаче извлечения экспериментальных данных экспериментов с моделью

Номер патента	GPT-4o-mini		GPT-4o	
	Количество извлеченных экспериментов/ Количество верно извлеченных экспериментов/ Фактическое количество эксперименто	Количество извлеченных таблиц/ Количество верно извлеченных таблиц/	Количество извлеченных экспериментов/ Количество верно извлеченных экспериментов/ Фактическое количество эксперименто	Количество извлеченных таблиц/ Количество верно извлеченных таблиц/
US6617405	2/0/3	3/0/6	3/3/3	5/4/6
US4481301	3/3/9	1/0/10	9/7/9	1/4/10
US93523408	10/0/13	1/0/6	10/7/13	3/4/6
US4849389	8/3/8	1/1/3	8/5/8	3/3/3
US4843132	3/3/3	1/0/1	3/3/3	1/1/1
Accuracy	**25%**	**4%**	**70%**	**61%**
1) При оценки точности не учитывались дубликаты, ошибки в ссылках, артефакты связзанные с ошибками в исходном текстовом слое				

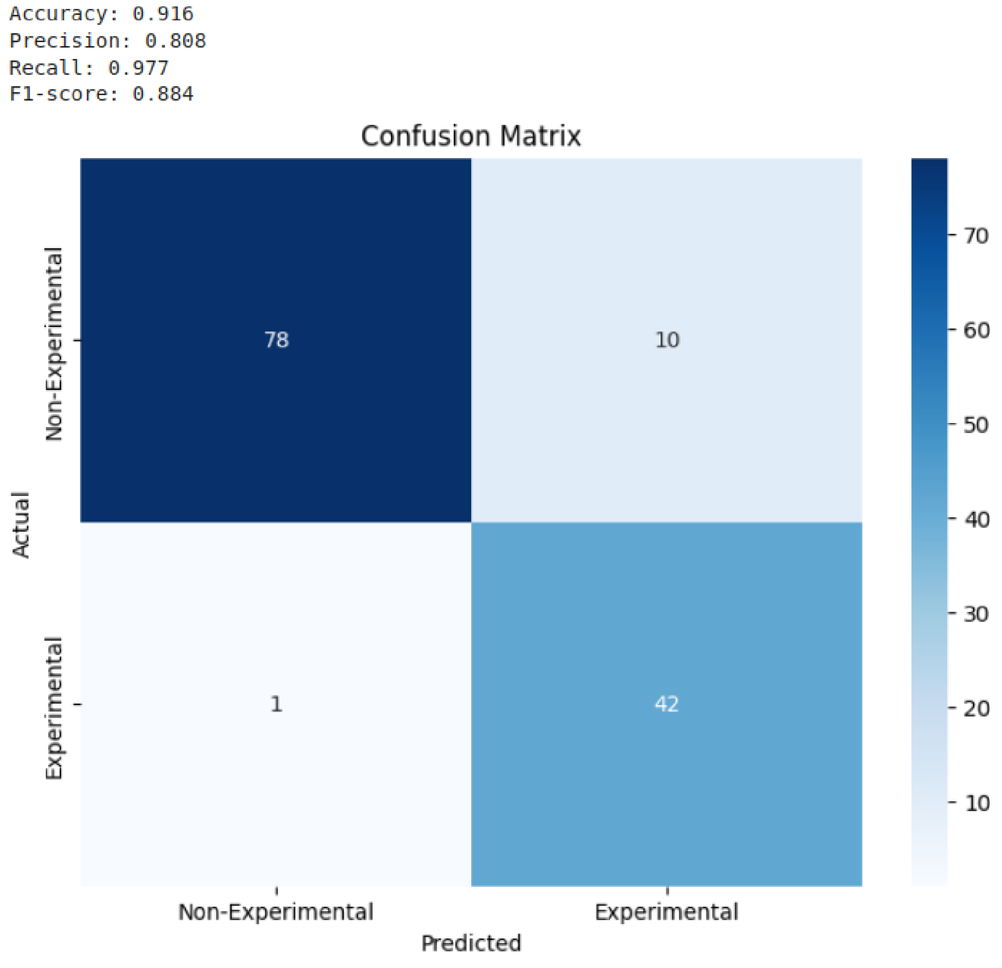
Анализа результатов работы модели в узком донемене требует трудозатратной ручной проверки. Это значительно усложняет разработку и настройку данных систем

Предварительная фильтрация текста

Один из вариантов для повышения качества — предварительная фильтрация текста для анализа. Для этого была протестирована быстрая и дешевая модель GPT-4o-mini.

Задача фильтрации ответить на вопрос: «Содержит ли данные текст релеватные экспреиментальные данные?».

Модель прекрасно справляется с классификацией текста. При анализе данных из 10 тестовых документов удалось добираться метрики recall 0.98. На выборке из более чем 100 страниц была допущена лишь одна FP незначительная ошикба. Данный подход позволяет сократить объем текста для анализе более чем в два раза.



Результаты классификации текста моделью GPT-4o-mini на страницы содержащие экспериментальные текст и не содержащие экспериментальные данные

Использование библиотеки docling

В 2024 году сотрудники компании IBM Research выложили в открытый доступ библиотеку Docling. Docling служит инструментов для обработки и парсинга различных форматов документов, включая PDF, DOCX, PPTX, HTML и изображения. Она преобразует эти форматы в единую представление, которое может быть использовано в приложениях генеративного ИИ, таких как RAG (Retrieve, Augment, Generate).

В частности библиотека представляет удобный инструментарий для использования разсличных технологий распознования текста неразмеченных PDF файлов и их преобразования в различные форматы с разметкой (Markdown, HTML)

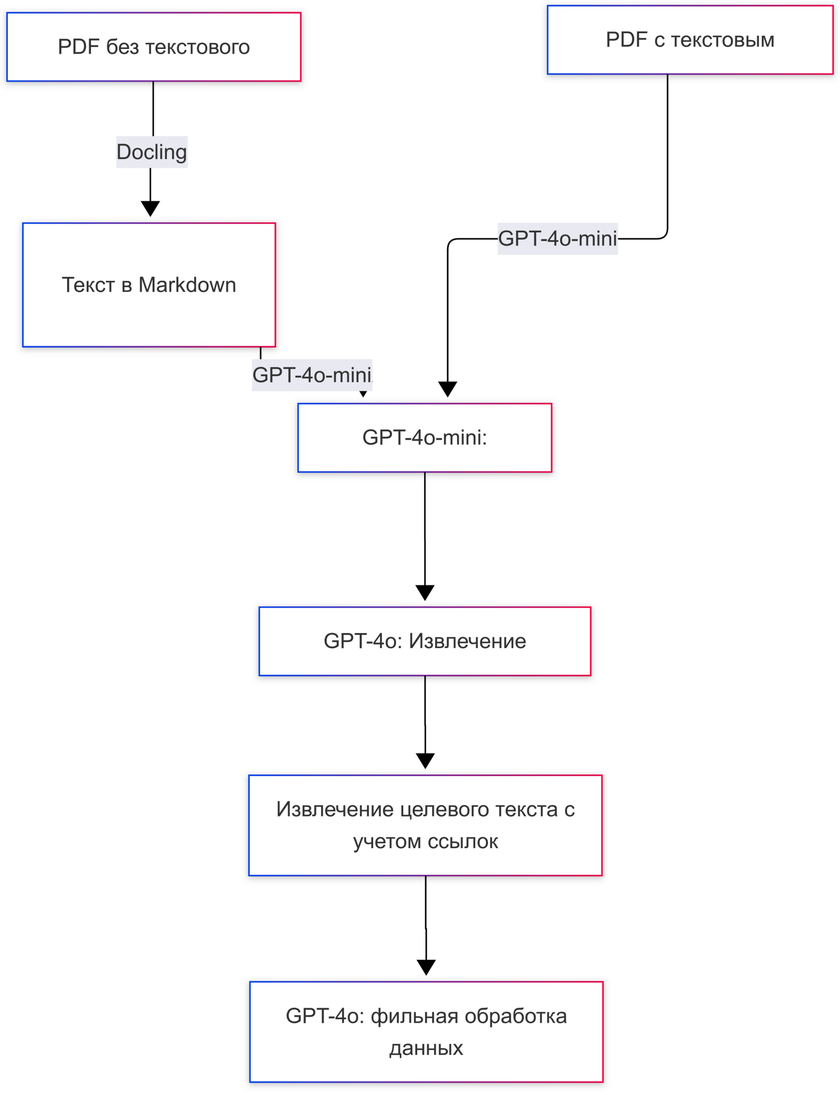
С помощью docling можно настроить качественную комбинированную обработку таблиц. На рисунке справа представлен пример такой обработки.

TABLE 1

EXAMPLE	13	14	15	16	17	18	19	20	21	22	23
Catalyst used	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7	Ex. 8	Ex. 9	Ex. 10	Ex. 10
Ti (%)	7.0	1.4	1.2	1.5	1.7	1.8	2.1	2.0	1.9	2.0	2.0
Mg (%)	2.0	0.5	0.3	0.3	2.4	2.7	1.3	1.5	2.9	1.5	1.5
Temperature (° C.)	88	88	88	88	88	88	88	88	88	88	75
Ethylene partial pressure (bar)	3	11.8	9	10.9	5	7	7	7	7	7	7
Total pressure (bar)	21	22	21	21	21	21	21	21	21	21	21
Residence Time (h)	4.43	3.71	5.32	3.20	4.86	3.92	3.18	2.89	3.91	3.70	3.15
H_2/ethylene (mole/mole)	0.10	0.098	0.086	0.096	0.096	0.118	0.098	0.12	0.13	0.12	0.17
Butene/ethylene (mole/mole)	0.39	0.41	0.46	0.44	0.35	0.40	0.39	0.41	0.37	0.41	0.45
Catalytic yield (kg/g)	3.0	1.6	1.7	1.6	4.3	7.5	5.0	7.5	7.1	7.1	6.2
Bulk Density (g/cm^3)	0.28	0.34	0.38	0.34	0.29	0.31	0.38	0.38	0.38	0.38	0.35
MIE (g/10 min)	0.68	0.66	0.72	0.74	0.64	0.68	0.69	0.67	0.69	0.65	0.57
MIF (g/10 min)	20.4	17.3	20.6	19.9	23.3	18.6	19.1	17.6	18.2	16.9	15.2
MFR - F/E	30	26	28	27	36	27	28	26	26	26	26
Comonomer Content (%)	8.8	7.8	8.7	7.7	9.4	8.3	8.4	7.9	8.2	8.1	8.8
Fraction Soluble in Xylene (%)	12.3	8.9	9.4	7.1	13.7	9.4	9.5	7.8	10.1	8.0	9.3
Polymer Density (g/cm^3)	0.918	0.918	0.918	0.919	0.917	0.918	0.918	0.918	0.918	0.918	0.917
Flowability (s/100 g)	—	14.4	11.1	13.3	14.4	17.9	12.4	11.9	10.8	12.7	12.6
Particle size distribution (wt %)											
<250 μm	2	2	1	1	2	3	<0.5	1	1	1	1
250-420 μm	7	7	7	2	7	5	2	2	2	2	2
420-840 μm	37	49	60	35	45	28	36	35	28	39	37
>840 μm	54	42	32	62	46	64	62	62	69	58	60
Haze (%)	11.3	—	—	—	—	—	11.6	11.3	11.3	11.4	—
Gloss (%)	73.0	—	—	—	—	—	79.5	80.7	79.3	80.2	—
Blocking (g/100 cm^2)	—	22	21	15	—	24	24	16	30	19	24

```
table_data = """[TableData(caption=None, headers=[], rows=[], reference=[]),
 TableData(caption='Table 1', headers=['EXAMPLE', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22', '23'],
 rows=[['Catalyst used', 'Ex. 1', 'Ex. 2', 'Ex. 3', 'Ex. 4', 'Ex. S', 'Ex. 6', 'Ex. 7', 'Ex. 8', 'Ex. 9', 'Ex. 10', 'Ex. 10'],
 ['Ti(%)', '7.0', '1.4', '1.2', '1.5', '1.7', '1.8', '2.1', '2.0', '1.9', '2.0', '2.0'],
 ['Mg (%)', '2.0', 'O.S', 'O.3', 'O.3', '2.4', '2.7', '1.3', '1.5', '2.9', '1.5', '1.5'],
 ['Temperature (C.)', '88', '88', '88', '88', '88', '88', '88', '88', '88', '88', '75'],
 ['Ethylene partial pressure (bar)', '3', '11.8', '9', '10.9', '5', '7', '7', '7', '7', '7', '7'],
 ['Total pressure (bar)', '21', '22', '21', '21', '21', '21', '21', '21', '21', '21', '21'],
 ['Residence Time (h)', '4.43', '3.71', '5.32', '3.20', '4.86', '3.92', '3.18', '2.89', '3.91', '3.70', '3.15'],
 ['Hafethylene (mole/mole)', '0.10', '0.098', '0.086', '0.096', '0.096', '0.118', '0.098', '0.12', '0.13', '0.12', '0.17'],
 ['Butenefethylene (mole/mole)', '0.39', '041', '046', '0.44', '0.35', '040', '0.39', '041', '0.37', '041', '045'],
 ['Catalytic yield (kg/g)', '3.0', '1.6', '1.7', '1.6', '4.3', '7.5', 'S.O', '7.5', '7.1', '7.1', '6.2'],
 ['Bulk Density (g/cm)', '0.28', '0.34', '0.38', '0.34', '0.29', '0.31', '0.38', '0.38', '0.38', '03S'], ['MIE (g/10 min)', '0.68', '0.66', '0.72', '0
 ['MIF (g/10 min)', '20.4', '17.3', '206', '19.9', '23.3', '18.6', '19.1', '17.6', '18.2', '16.9', '15.2'], ['MFR - FE', '30', '26', '28', '27', '36', '27', '
 ['Comonomer Content (%)', '8.8', '7.8', '8.7', '7.7', '9.4', '8.3', '8.4', '7.9', '8.2', '8.1', '8.8'], ['Fraction Soluble in Xylene (%)', '12.3', '8.9', '9.
 ['Polymer Density (g/cm)', '0.918', '0.918', '0.918', '0.919', '0.917', '0.918', '0.918', '0.918', '0.918', '0.917'], ['Flowability (s/100g)', '', '
 ['Particle size distribution (wt %)', '', '', '', '', '', '', '', '', '', '', ''],
 ['<250 m', '2', '2', '1', '1', '2', '3', '<0.S', '1', '1', '1', '1'],
 ['250-420 m', '7', '7', '7', '2', '7', '5', '2', '2', '2', '2', '2'],
 ['420-840 m', '37', '49', '60', '35', '45', '28', '36', '35', '28', '39', '37'],
 ['>840 m', 'S4', '42', '32', '62', '46', '64', '62', '62', '69', '58', '60'],
 ['Haze (%)', '11.3', '', '', '', '', '', '11.6', '11.3', '11.3', '11.4', ''], ['Gloss (%)', '73.0', '', '', '', '', '', '79.5', '80.7', '79.3', '80.2', ''],
 ['Blocking (g/100 cm)', '', '22', '21', '15', '', '24', '24', '16', '30', '19', '24']],
 reference=['Examples 1', 'Examples 2', 'Examples 3', 'Examples 4', 'Examples 5', 'Examples 6', 'Examples 7', 'Examples 8', 'Examples 9', 'Examples 10'])]"""
```

Алгоритм работы



Пример извлеченных данных

```python
class UnitValue(BaseModel):…

class ReactionType(str, Enum):…

class MonomerType(str, Enum):…

class PolymerProperties(BaseModel):
    mw: Optional[UnitValue] = Field(None, description="Weight average molecular weight")
    mn: Optional[UnitValue] = Field(None, description="Number average molecular weight")
    pdi: Optional[float] = Field(None, description="Polydispersity index")
    density: Optional[UnitValue] = Field(None, description="Density")
    tm: Optional[UnitValue] = Field(None, description="Melting temperature")
    bulk_density: Optional[UnitValue] = Field(None, description="Bulk density")
    melt_index: Optional[UnitValue] = Field(None, description="Melt index")
    melt_flow_rate: Optional[UnitValue] = Field(None, description="Melt flow rate")

class PolytestResult(BaseModel):
    reaction_type: Optional[ReactionType] = Field(None, description="Type of reaction in the polytest")
    monomer_type: Optional[MonomerType] = Field(None, description="Type of monomer used in the polytest")
    comonomer_type: Optional[MonomerType] = Field(None, description="Type of comonomer if used")
    reactor_volume: Optional[UnitValue] = Field(None, description="Volume of reactor")
    temperature: Optional[UnitValue] = Field(None, description="Temperature of polytest")
    monomer: Optional[UnitValue] = Field(None, description="Pressure of monomer")
    hydrogen: Optional[UnitValue] = Field(None, description="Amount of hydrogen used")
    catalyst_amount: Optional[UnitValue] = Field(None, description="Amount of catalyst")
    cocatalyst_amount: Optional[UnitValue] = Field(None, description="Amount of cocatalyst")
    reaction_time: Optional[UnitValue] = Field(None, description="Reaction time")
    activity: Optional[UnitValue] = Field(None, description="Catalyst activity")
    polymer_properties: Optional[PolymerProperties] = Field(None, description="Polymer properties")

class Component(BaseModel):
    name: Optional[str] = Field(None, description="Name of the component")
    amount: Optional[UnitValue] = Field(None, description="Amount of component")

class SynthesisConditions(BaseModel):
    temperature: Optional[UnitValue] = Field(None, description="Temperature in °C, if applicable")
    time: Optional[UnitValue] = Field(None, description="Duration, if applicable")
    pressure: Optional[UnitValue] = Field(None, description="Pressure in MPa if applicable")

class SynthesisStep(BaseModel):
    step_number: Optional[int] = Field(None, description="Sequential number of the synthesis step")
    description: Optional[str] = Field(None, description="Description of the step")
    components_added: Optional[List[Component]] = Field(None, description="Components added in this step")
    conditions: Optional[SynthesisConditions] = Field(None, description="Conditions during this step")

class CatalystSynthesis(BaseModel):
    id: str = Field(description="Unique identifier for the catalyst synthesis")
    synthesis_steps: Optional[List[SynthesisStep]] = Field(None, description="Ordered list of synthesis steps")
    polytest_results: Optional[List[PolytestResult]] = Field(None, description="Results of polytests if applicable")
    catalyst_composition: Optional[List[Component]] = Field(None, description="Composition of final product")
```

[ExperimentInfo(

id='Example **1**',

text_of_example='In a 5 liter flask fitted with a mechanical stirrer and previously purged with nitrogen were fed 44 g (0.462 moles) of anhydrous MgCl, and 330 ml (0.969 moles) of Ti(OBu). This mixture was allowed to stir at 300 rpm and heated to 150°C. for about 12 hours in order to have the solids completely dissolved, thereby a clear liquid product was obtained. This resulting liquid was cooled down to 40°C. and under gently stirring at 150 rpm, it was diluted with 3200 ml of anhydrous hexane. Into this solution kept at 40° C. and under the same stirring, 250g of the silica support were added. This silica was previously dehydrated and treated with 19 ml (0.139 moles) of triethylaluminum diluted in anhydrous hexane, for 50 minutes and at room temperature. Once the addition of the silica is completed, the mixture was heated to 60° C. and kept at this temperature for 1 hour. Into this mixture, kept at 60° C. and under gently stirring, a solution consisting of 100 ml of anhydrous hexane and 192 ml of PMHS (0.085 moles) was dropped into it over a period of time of 1.5 hours. At the end of the addition, stirring was continued for 2 hours at a temperature of 60° C. To this mixture a solution of 200 ml of anhydrous hexane and 184 ml of SiC1 (1.606 moles) was dropped over a period of time of 1 hour. At the end of the addition, stirring was continued for 3.5 hours at a temperature of 60°C. The temperature of the mixture was then brought to 65° C. and kept for additional 2 hours. After cooling the mixture to room temperature, the stirring was stopped to have the solid settled. The supernatant liquid was removed, the solid was repeatedly washed with anhydrous hexane and then dried at 60° C. under nitrogen flow thus giving 390 g of a reddish powder.\n\nThe chemical and physical characteristics of the resulting reddish powder were as follows:\n\nTotal Titanium-7.0% (by weight)\nMg 2.0% (by weight)\nSiO=75.9% (by weight)\nAl=0.5% (by weight)\nCl=10.9% (by weight)\nOBu=4.1% (by weight)\nSurface Area (B.E.T.)=200 m/g\nPore Volume (B.E.T.)=0.45 cm/g',

type='catalyst_synthesis',

reference=[]),

Пример извлеченных данных

[ExperimentInfo(id='Example 1', text_of_example='In a 5 liter flask fitted with a mechanical stirrer and previously purged with nitrogen were fed 44 g (0.462 moles) of anhydrous MgCl, and 330 ml (0.969 moles) of Ti(OBu). This mixture was allowed to stir at 300 rpm and heated to 150°C. for about 12 hours in order to have the solids completely dissolved, thereby a clear liquid product was obtained. This resulting liquid was cooled down to 40°C. and under gently stirring at 150 rpm, it was diluted with 3200 ml of anhydrous hexane. Into this solution kept at 40° C. and under the same stirring, 250g of the silica support were added. This silica was previously dehydrated and treated with 19 ml (0.139 moles) of triethylaluminum diluted in anhydrous hexane, for 50 minutes and at room temperature. Once the addition of the silica is completed, the mixture was heated to 60° C. and kept at this temperature for 1 hour. Into this mixture, kept at 60° C. and under gently stirring, a solution consisting of 100 ml of anhydrous hexane and 192 ml of PMHS (0.085 moles) was dropped into it over a period of time of 1.5 hours. At the end of the addition, stirring was continued for 2 hours at a temperature of 60° C. To this mixture a solution of 200 ml of anhydrous hexane and 184 ml of SiC1 (1.606 moles) was dropped over a period of time of 1 hour. At the end of the addition, stirring was continued for 3.5 hours at a temperature of 60°C. The temperature of the mixture was then brought to 65° C. and kept for additional 2 hours. After cooling the mixture to room temperature, the stirring was stopped to have the solid settled. The supernatant liquid was removed, the solid was repeatedly washed with anhydrous hexane and then dried at 60° C. under nitrogen flow thus giving 390 g of a reddish powder.\n\nThe chemical and physical characteristics of the resulting reddish powder were as follows:\n\nTotal Titanium-7.0% (by weight)\nMg 2.0% (by weight)\nSiO=75.9% (by weight)\nAl=0.5% (by weight)\nCl=10.9% (by weight)\nOBu=4.1% (by weight)\nSurface Area (B.E.T.)=200 m/g\nPore Volume (B.E.T.)=0.45 cm/g', type='catalyst_synthesis', reference=[]),

```
play(synthesis_df)\n    \n    print("\nPolytest Results DataFrame:")\n    display(polytest_df)\n'
```

[98]:
```
synthesis_steps_df, polytest_results_df = process_catalyst_synthesis(result)

print("Synthesis Steps:")

synthesis_steps_df
```

[98]:
Synthesis Steps:

	step_number	description	temperature_value	temperature_unit	time_value	time_unit	component_name	component_amount_value	component_amount_unit
0	1	Feed anhydrous MgCl and Ti(OBu) into a flask f...	150.0	°C	12.0	h	anhydrous MgCl	44.0	g
1	1	Feed anhydrous MgCl and Ti(OBu) into a flask f...	150.0	°C	12.0	h	Ti(OBu)	330.0	ml
2	2	Cool the mixture to 40°C and dilute with anhyd...	40.0	°C	NaN	NaN	anhydrous hexane	3200.0	ml
3	3	Add dehydrated silica support treated with tri...	40.0	°C	50.0	min	silica support	250.0	g
4	3	Add dehydrated silica support treated with tri...	40.0	°C	50.0	min	triethylaluminum	19.0	ml
5	4	Heat the mixture to 60°C and add a solution of...	60.0	°C	1.5	h	anhydrous hexane	100.0	ml

Отфильтрованный и подготвленный текст с высокой степенью точностью переводится в структурированный тип данных.
Полученные структурированные данные можно обрабатывать стандартными методами анализа данных.

Выводы

Проведено исследование применимости передовых LLM для анализа текстовой информации в домене «катализаторы полимеризации олефинов»:

1) Показано, что модели семейства gpt-4o способны классифицировать сложный научный текст с целью выяляния с высокой точностью (значение целовой метрики recall 0.98);

2) Подобрана конфигурация, позволяющая проводить сложную аугментацию исходного текст учитывающего внутренние взаимосвязи отдельных частей документа;

3) Показана возможность извлечения с выской точностью данных из аугментированного текста в сложные структуры данных с выской степенью «вложенности»

Дальнейшие шаги

1) На основе полученного в ходе эксперментов подхода необходимо реализовать приложение для удобной обработки документов
2) Провести ручную оценку качества извлечения на выборке из 10 документов

СПИСОК ЛИТЕРАТУРЫ

1. Science of science | Science [Electronic resource]. URL: https://www.science.org/doi/10.1126/science.aao0185 (accessed: 01.11.2024).

2. Ananikov V. Top 20 Influential AI-Based Technologies in Chemistry. Chemistry, 2024.

3. Ramos M.C., Collison C.J., White A.D. A Review of Large Language Models and Autonomous Agents in Chemistry: arXiv:2407.01603. arXiv, 2024.

4. Lála J. et al. PaperQA: Retrieval-Augmented Generative Agent for Scientific Research: arXiv:2312.07559. arXiv, 2023.

5. Zheng Z. et al. ChatGPT Chemistry Assistant for Text Mining and Prediction of MOF Synthesis.