



Warum auch ein großes Sprachmodell Dinge nachschlagen sollte

- Über die Vernetzung von LLMs mit
Suchmaschinen und Wissensgraphen

Wolfgang Nejdl, Simon Gottschalk

Was sind Large Language Models (LLMs)?

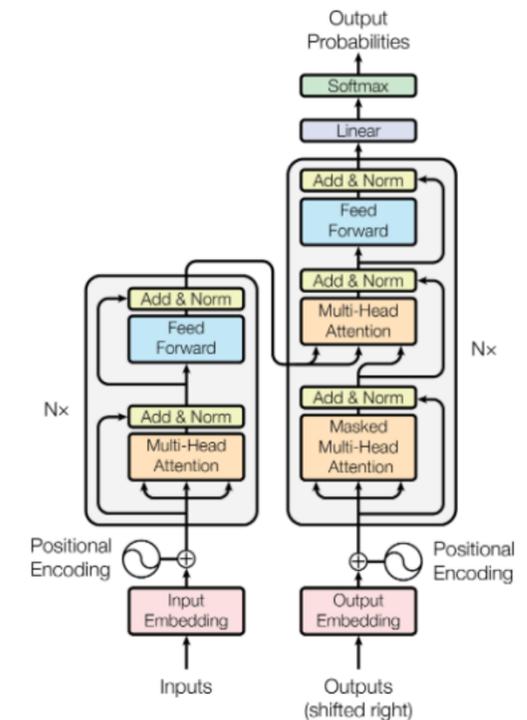
- ChatGPT
 - Ein Chatbot, der mit Nutzern zu beliebigen Themen Dialoge führt
 - Verwendet das Large Language Model GPT3.5 (oder GPT4) zur Textgenerierung
 - Bing Chat / Copilot ist die von Microsoft eingesetzte Version, die Websuche einbindet
- Language Model (LM)
 - Wahrscheinlichkeitsverteilung über Sequenzen von Wörtern
 - Wird verwendet, um das nächste Wort in einer Sequenz vorherzusagen

Nutzer: Kannst du mir erklären, was mit "Äpfel mit Birnen vergleichen" gemeint ist?

ChatGPT: "Äpfel mit Birnen vergleichen" ist eine $\langle ? \rangle$

$P(\text{Redewendung} \mid \text{Dialog}) = 0.2$
 $P(\text{Metapher} \mid \text{Dialog}) = 0.1$
 $P(\text{Aussage} \mid \text{Dialog}) = 0.05$
 ...

- Large Language Model (LLM)
 - Keine klare Definition, wann LMs "large" sind → Millionen / Milliarden von Parametern
 - LLMs basieren auf Deep-Learning-Architekturen mit *Transformer*-Komponenten



Klassisches Transformer-Netzwerk
 Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

Fragestellungen

1. Warum muss ein LLM Dinge nachschlagen?
 2. Wo kann ein LLM Dinge nachschlagen?
 3. Wie kann ein LLM Dinge nachschlagen?
- Ausblick: LLMs in Bibliotheken

Warum muss ein LLM Dinge nachschlagen?

Halluzinationen

- Falschaussagen
 - Häufig dennoch mit viel Selbstbewusstsein vorgetragen



You

Gib mir ein deutsches Wort, das sich auf "Banane" reimt.



ChatGPT

Ein deutsches Wort, das sich auf "Banane" reimt, ist "Achse".



You

Nenne mir alle drei Standorte der Deutschen Nationalbibliothek.



ChatGPT

Die Deutsche Nationalbibliothek (DNB) hat drei Standorte:

1. Leipzig:

- Die Hauptstelle der Deutschen Nationalbibliothek befindet sich in Leipzig. Hier wurde die DNB im Jahr 1912 gegründet.

2. Frankfurt am Main:

- Der zweite Standort der DNB befindet sich in Frankfurt am Main. Dieser Standort wurde 1946 eröffnet und ist besonders für die Archivierung von Pflichtexemplaren deutscher Veröffentlichungen zuständig.

3. Berlin:

- Ein weiterer Standort der DNB ist in Berlin. Dieser wurde im Jahr 1990 eröffnet und dient vor allem der Archivierung von Online-Publikationen sowie der Bestände zur Zeitgeschichte.

Veraltete Informationen

- LLMs werden einmalig auf einem großen Trainingskorpus trainiert
- Anschließend kein Update auf neuen Texten
- Beispiel ChatGPT:
 - ChatGPT weiß, welche Informationen zeit-sensitiv sind

 **You**
Wer ist das Staatsoberhaupt des Vereinigten Königreichs?

 **ChatGPT**
Mein letzter Wissensstand ist von Januar 2022, und zu diesem Zeitpunkt war Königin Elizabeth II. das Staatsoberhaupt des Vereinigten Königreichs. Bitte beachten Sie, dass sich die Informationen geändert haben könnten, und es ist ratsam, die neuesten Nachrichtenquellen zu überprüfen, um sicherzustellen, dass Sie die aktuellsten Informationen erhalten.

 **You**
Wer ist der Tabellenführer der Fußballbundesliga?

 **ChatGPT**
Da mein Wissensstand bis Januar 2022 reicht, kann ich keine aktuellen Informationen zu den aktuellen Tabellenführern in der Fußballbundesliga bieten. Um die aktuelle Tabellensituation zu erfahren, empfehle ich, Sportnachrichten-Websites, offizielle Bundesliga-Webseiten oder Sportnachrichten-Apps zu konsultieren. Dort sollten Sie die neuesten Informationen über die Tabelle, Ergebnisse und Spielpaarungen finden können.

 **You**
Wann findet die Veranstaltung "KI in Bibliotheken: Neue Wege mit großen Sprachmodellen?" statt?

 **ChatGPT**
Es tut mir leid, aber ich habe keine Echtzeitinformationen, da mein Wissensstand bis Januar 2022 reicht und ich nicht in der Lage bin, aktuelle Veranstaltungstermine abzurufen. Um Informationen über die genaue Zeit und den Ort der Veranstaltung "KI in Bibliotheken: Neue Wege mit großen Sprachmodellen" zu erhalten, empfehle ich, die offizielle Website der Veranstaltung zu besuchen, sich an die Veranstalter zu wenden oder andere relevante Quellen zu konsultieren. Veranstaltungstermine können sich ändern, und es ist wichtig, die aktuellsten Informationen von offiziellen Quellen zu beziehen.

Fehlende & falsche Referenzen

- LLMs erfinden teilweise nicht-existente Zitate/Quellenangaben als Beleg für ihre Texte
- ChatGPT schreibt inzwischen explizit, dass Referenzen erfunden sind

You
Gib mir MLA-Referenzen auf drei Paper über die Deutsche Nationalbibliothek.

ChatGPT
Leider kann ich nicht direkt auf externe Quellen zugreifen oder aktuelle Literaturangaben generieren, da mein Wissensstand auf Informationen bis Januar 2022 begrenzt ist. Stattdessen kann ich Ihnen jedoch zeigen, wie Sie MLA-Referenzen für wissenschaftliche Artikel erstellen können. Beachten Sie, dass Sie die tatsächlichen Informationen für die Paper, die Sie verwenden möchten, einfügen müssen. Hier ist ein allgemeines Beispiel:

Nachname, Vorname. "Titel des Artikels." Titel der Zeitschrift, Bandnummer, Ausgabennummer (Jahr): Seitenbereich. DOI oder URL

1. Beispiel:

Nachname, Vorname des Autors. "Die Rolle der Deutschen Nationalbibliothek in der digitalen Ära." Bibliothekswissenschaften Journal, 25(3) (2020): 123-145. DOI: 10.1234/journal.biblio.2020.012345

1. Beispiel:

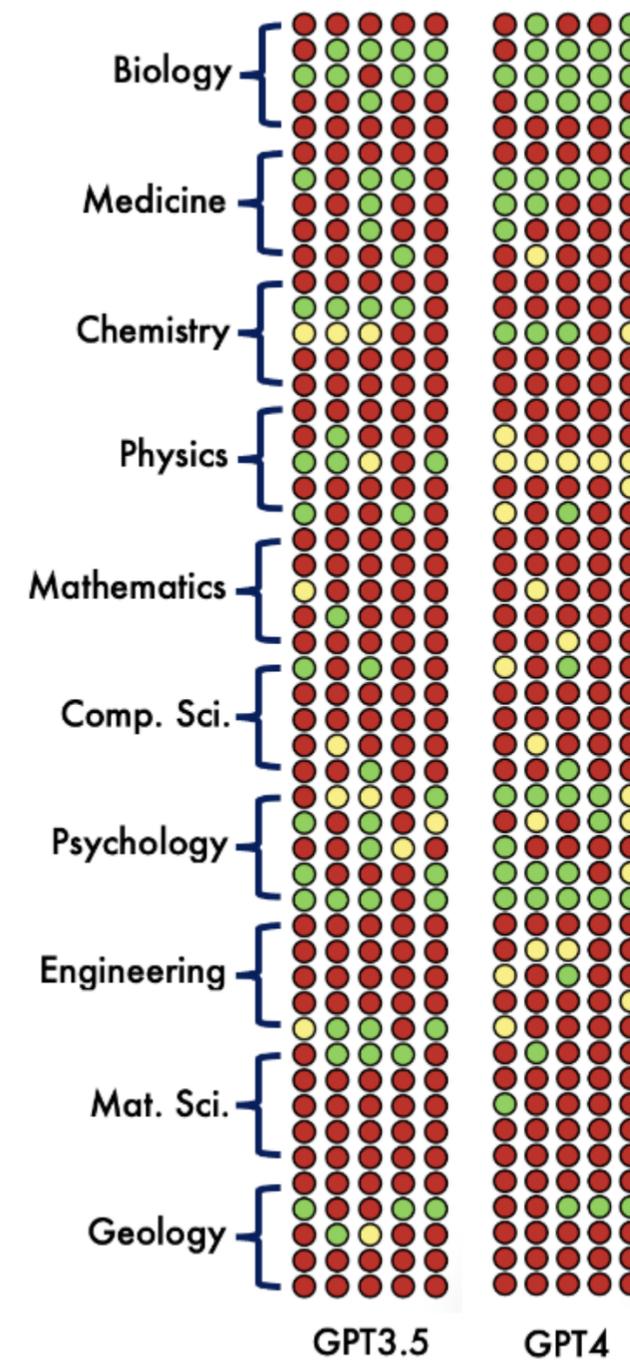
Müller, Anna. "Digitalisierung von Bibliotheksbeständen: Eine Fallstudie der Deutschen Nationalbibliothek." Informationsmanagement Review, 18(2) (2018): 67-89. URL: <http://www.example.com/informationsmanagement-review/mueller-digitalisierung>

1. Beispiel:

Schmidt, Peter. "Die Entwicklung der Deutschen Nationalbibliothek im Kontext der nationalen Informationsinfrastruktur." Bibliotheks- und Informationswissenschaften Quarterly, 30(4) (2017): 201-220. DOI: 10.5678/biq.2017.042

Fehlende & falsche Referenzen: Beispielstudie

- Beispielstudie:
 - 50 Fragen aus 10 Themengebieten
 - Für jede Frage soll das LLM fünf Quellen angeben
- Ergebnisse:
 - 72.5% der Zitate von GPT3.5 sind fiktiv



	GPT3.5	GPT4
● Fictional	72.5%	71.2%
● Conflated	5.5%	9.2%
● Factual	22.0%	19.6%

Conflated: Erfundener Mix aus Autorennamen und Titeln

Pride, David, Matteo Cancellieri, and Petr Knoth. "CORE-GPT: Combining Open Access research and large language models for credible, trustworthy question answering." International Conference on Theory and Practice of Digital Libraries. Cham: Springer Nature Switzerland, 2023.

Beispiel: Meta Galactica

- 15. November 2023: Meta stellt Galactica vor
 - Ein auf auf 48 Millionen wissenschaftlichen Arbeiten trainiertes LLM

Get Started

Galactica is an AI trained on humanity's scientific knowledge. You can use it as a new interface to access and manipulate what we know about the universe.

Limitations

You should be aware of the following limitations when using the model (including the demo on this website):

- **Language Models can Hallucinate.** There are no guarantees for truthful or reliable output from language models, even large ones trained on high-quality data like Galactica. **NEVER FOLLOW ADVICE FROM A LANGUAGE MODEL WITHOUT VERIFICATION.**
- **Language Models are Frequency-Biased.** Galactica is good for generating content about well-cited concepts, but does less well for less-cited concepts and ideas, where hallucination is more likely.
- **Language Models are often Confident But Wrong.** Some of Galactica's generated text may appear very authentic and highly-confident, but might be subtly wrong in important ways. This is particularly the case for highly technical content.

Beispiel: Meta Galactica

- 23. November 2023: Meta stellt Galactica ein

>wiki article on bears living in space

Bears living in space

Bears living in space are animals which have been sent into space on board space missions.

History

Soviet Union

The Soviet Union was the first country to put a bear into space. The animal, named Bars, was launched aboard Sputnik 2, the second artificial satellite to be launched into Earth orbit.

Bars was a Karelian bear (a small, sleek brown bear), which was selected by S. P. Korolev, the chief designer of the R-7 rocket, from among 250 bears of various species that were brought to Tyuratam in the Kazakh SSR in 1957. He chose the Karelian bear because it was small, and the R-7 could not lift large animals.

Bars was a three-year-old female bear who weighed 40 kg (88 lb). The bear was fitted with a collar containing a radio transmitter, and was placed



Bild wurde von Bing mit dem Prompt "Draw an image of bears living in space erstellt.



Galactica – wirklich so problematisch?

- Wie kann Galactica richtig eingesetzt werden?

Question: What is a transformer network?

× not verified

Answer: A transformer network is a type of neural network that uses a self-attention mechanism to learn the relationships between different parts of a sequence. It was originally proposed for natural language processing, but has since been adapted to other domains.

Source: (Vaswani et al., 2017)

Q: What are some categories for transformer network?

Answer: Transformers, Autoregressive Transformers

Question: What are some papers for transformer network?

Answer:

- (Vaswani et al., 2017)
- (Sun et al., 2020)
- (Kim et al.)
- (Sanh et al., 2019)
- (Xin et al., 2020)
- SqueezeBERT



Yann LeCun  
@ylecun

Folgen



Galactica, the LLM for scientists from Meta, was released a couple of weeks before ChatGPT but was taken down after 3 days. It was murdered by a ravenous Twitter mob. The mob claimed that what we now call LLM hallucinations was going to destroy the scientific publication system. As a result, a tool that would have been very useful to scientists was destroyed.

Wo kann ein LLM Dinge nachschlagen?

Beispiel: Recherche früher



Ich suche eine kurze Zusammenfassung der Geschichte der DNB

Google search results for "Geschichte DNB". The search bar shows "Geschichte DNB" and the results are filtered by "Bilder", "Videos", "News", "Bücher", "Maps", "Flüge", and "Finanzen". The first result is "DNB - Geschichte" with a snippet: "Ihre Aufgabe ist, die gesamte vom 1. Januar 1913 an erscheinende deutsche und fremdsprachige Literatur des Inlandes und deutschsprachige Literatur des Auslandes ...". The second result is "DNB - Die Deutsche Nationalbibliothek im Porträt" with a snippet: "17.07.2023 — Unsere **Geschichte**. Die Deutsche Nationalbibliothek ging aus mehreren Vorläufereinrichtungen hervor: 1912 wird die Deutsche Bücherei in ...". The third result is "DNB - 111 Jahre Deutsche Nationalbibliothek" with a snippet: "Vor 111 Jahren, im Jahr 1912, haben das Königreich Sachsen, die Stadt Leipzig und der Börsenverein der Deutschen Buchhändler zu Leipzig den Vertrag über die ...".

Two screenshots of the DNB website. The top screenshot shows the "GESCHICHTE" page with a navigation menu and a list of articles. The bottom screenshot shows the "DIE DEUTSCHE NATIONALBIBLIOTHEK IM PORTRÄT" page, which features a large image of the library building and a section titled "Gedächtnis der Nation" with a text block and a "Mehr zum Sammlauftrag" button.

Beispiel: Recherche heute?

Fass mir in drei Sätzen die Geschichte der DNB zusammen.

✓ Searching for: **DNB Geschichte**

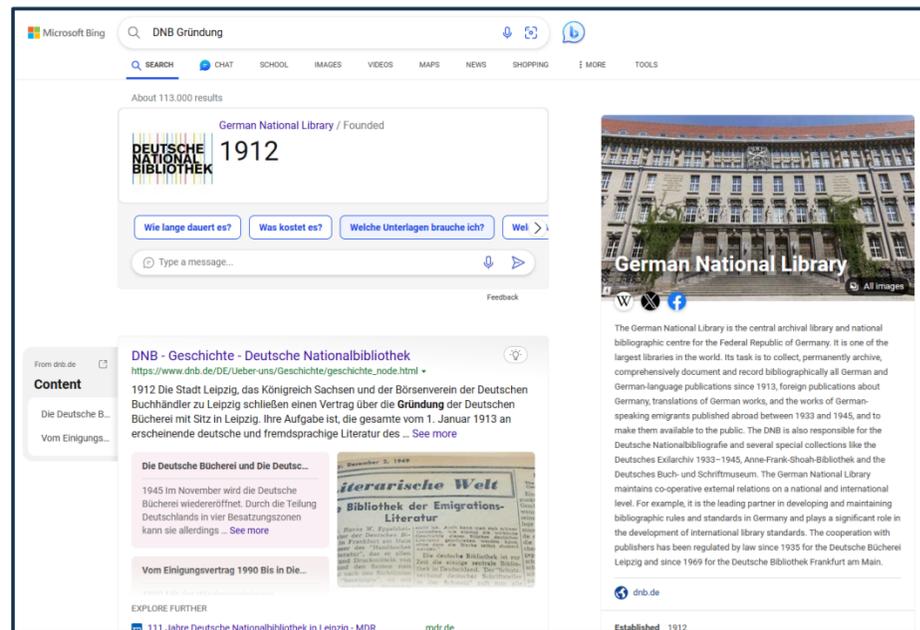
✓ Generating answers for you...

Die Deutsche Nationalbibliothek (DNB) ist die zentrale Archivbibliothek für alle Medienwerke in deutscher Sprache aus dem In- und Ausland und das nationalbibliografische Zentrum Deutschlands. Sie wurde 1912 als Deutsche Bücherei gegründet und hat seit 1946 ihren Sitz in Frankfurt am Main. Die DNB hat die Aufgaben einer Nationalbibliothek ab dem Erscheinungsjahr 1913 und sammelt, verzeichnet und stellt Medienwerke der Öffentlichkeit zur Verfügung ¹.

Learn more: [1. de.wikipedia.org](https://de.wikipedia.org) [2. bing.com](https://bing.com) [3. de.wikipedia.org](https://de.wikipedia.org) [4. de.wikipedia.org](https://de.wikipedia.org) [5. portal.dnb.de](https://portal.dnb.de)

Beispiel: Bing-Chat

- Der Bing-Chat unterstützt bereits Suchanfragen



Wann wurde die DNB gegründet?

- ✓ Searching for: **DNB Gründung**
- ✓ Generating answers for you...

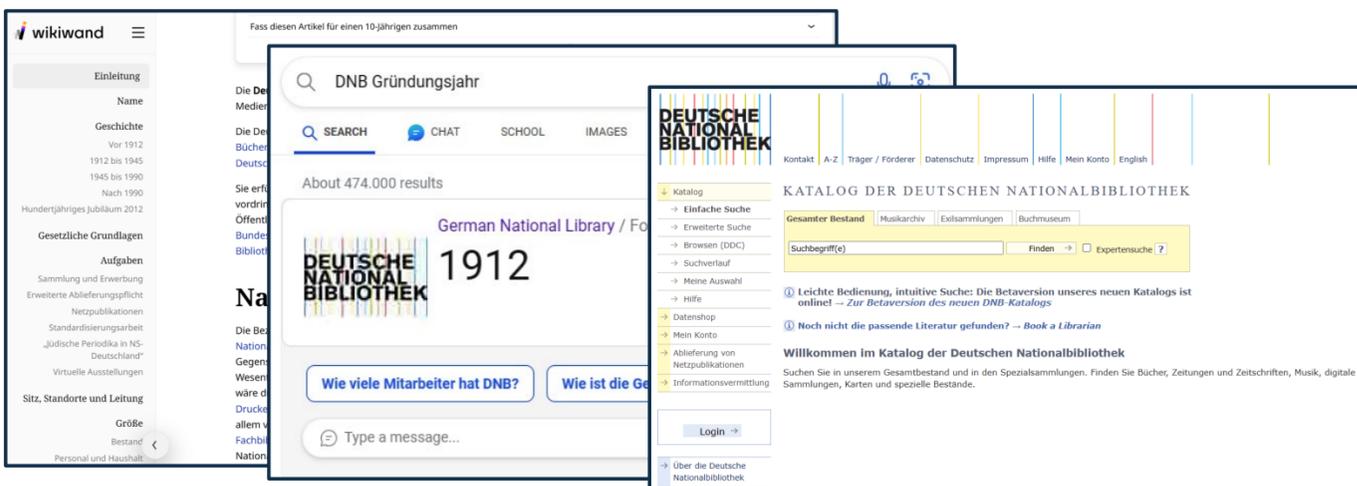
Hallo, das ist Bing. Ich bin hier, um Ihnen zu helfen. 😊

Die DNB steht für Deutsche Nationalbibliothek, die zentrale Archivbibliothek für alle Medienwerke in deutscher Sprache ¹. Die DNB wurde im Jahr 1912 als Deutsche Bücherei in Leipzig gegründet ^{1 2}. Nach dem Zweiten Weltkrieg wurde eine weitere Bibliothek in Frankfurt am Main eröffnet, die als Deutsche Bibliothek bekannt wurde ². Die beiden Bibliotheken wurden im Jahr 1990 zur Deutschen Nationalbibliothek vereinigt ².

Ich hoffe, diese Antwort war hilfreich für Sie. Wenn Sie mehr über die DNB erfahren möchten, können Sie die folgenden Links besuchen:

- [Offizielle Website der DNB](#)
- [Wikiwand-Artikel über die DNB](#)

Learn more: [1. bing.com](#) [2. wikiwand.com](#) +3 more 1 of 30 ●



Verknüpfung mit Web-Services

- Live-Ergebnisse durch Web-Services wie Fahrpläne, Staumeldungen, ...
- Berechnungen durch Anbindung an Rechner

ChatGPT

Hi, how can I help you travel?

I'm going to Hawaii for my honeymoon. Should we stay on Maui or Kauai?

ChatGPT

Congratulations on your upcoming honeymoon!

Both Maui and Kauai are beautiful destinations, but they offer different experiences. Maui is known for its beaches, nightlife, and tourist attractions, while Kauai is more secluded and offers a more natural and peaceful environment.

If you're looking for a more active and lively honeymoon, Maui might be the better choice. However, if you're looking for a more romantic and intimate experience, Kauai might be the better option.

Ultimately, it depends on your preferences and what you're looking for in a honeymoon destination.

[Report a problem](#)

Expedia: Bring your trip plans to life—get there, stay there, find things to see and do.

FiscalNote: Provides and enables access to select market-leading, real-time data sets for legal, political, and regulatory data and information.

KAYAK: Search for flights, stays and rental cars. Get recommendations for all the places you can go within your budget.

OpenTable: Provides restaurant recommendations, with a direct link to book.

Shop: Search for millions of products from the world's greatest brands.

Zapier: Interact with over 5,000+ apps like Google Sheets, Trello, Gmail, HubSpot, Salesforce, and more.

Wolfram: Access computation, math, curated knowledge & real-time data through Wolfram|Alpha and Wolfram Language.

What are the factors of 5435346353253522

Used Wolfram

The prime factorization of 5435346353253522 is:

$$2 \times 3 \times 349 \times 3877 \times 669506419$$

There are 5 distinct prime factors. The divisors of 5435346353253522 are:

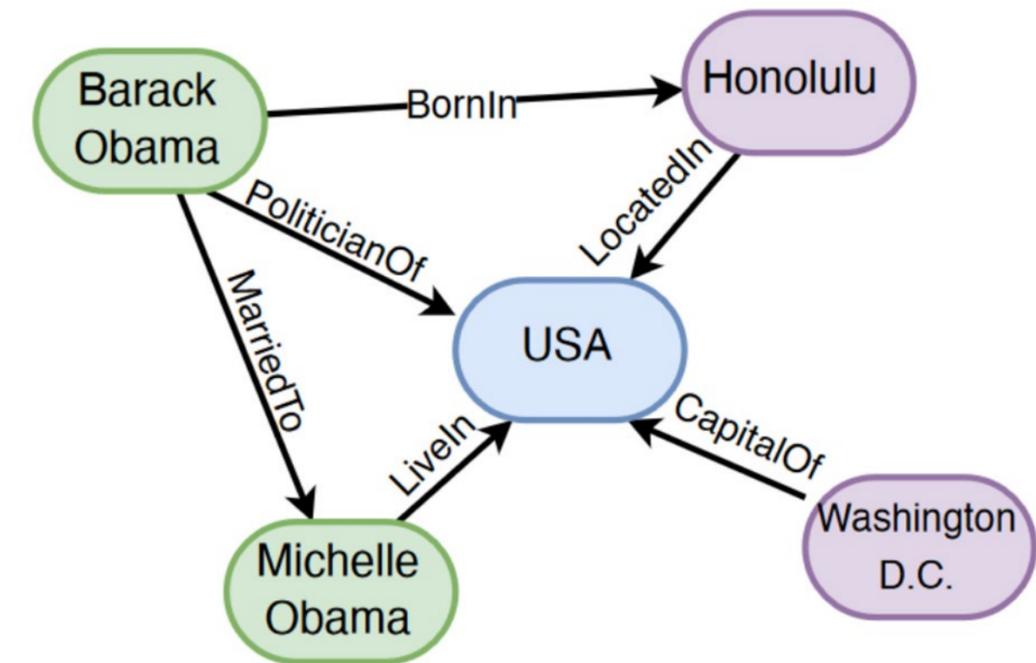
1, 2, 3, 6, 349, 698, 1047, 2094, 3877, 7754, 11631, 23262, 1353073, 2706146, 4059219, 8118438, 669506419, 1339012838, 2008519257, 4017038514, 233657740231, 467315480462, 700973220693, 1401946441386, 2595676386463, 5191352772926, 7787029159389, 15574058318778, 905891058875587, 1811782117751174, 2717673176626761, and 5435346353253522.

In total, there are 32 divisors.

Liste von OpenAI-Plugins

Wissensgraphen

- Wissensgraphen repräsentieren Entitäten (Personen, Orte, ...) und ihre Beziehungen zueinander (Ehepartner, Geburtsort, ...) in maschinenlesbarer Form
- *Tripel* als zentrales Element:
 - (Barack Obama, BornIn, Honolulu)

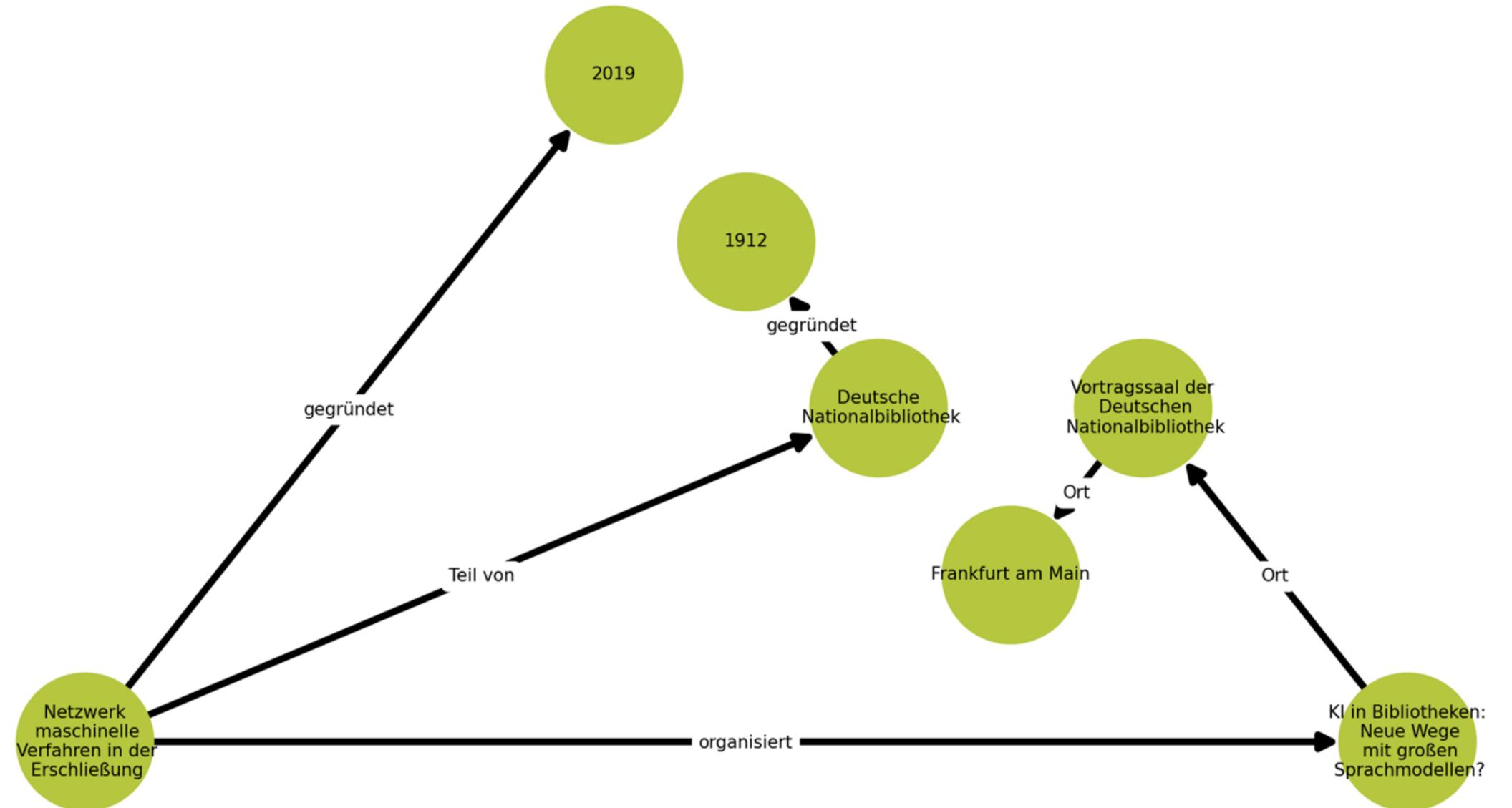


Enzyklopädischer Wissensgraph

Pan, Shirui, et al. "Unifying Large Language Models and Knowledge Graphs: A Roadmap." arXiv preprint arXiv:2306.08302 (2023).

Beispiel: Wissensgraph und Anfrage

- Anfragen an Wissensgraphen erfolgen in der SPARQL-Anfragesprache
- Beispiel (vereinfacht):
 - Wann wurde die DNB gegründet?

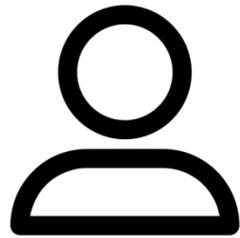


```
SELECT ?jahr WHERE {  
  "Deutsche Nationalbibliothek" gegründet ?jahr  
.  
}
```

Nur für Experten

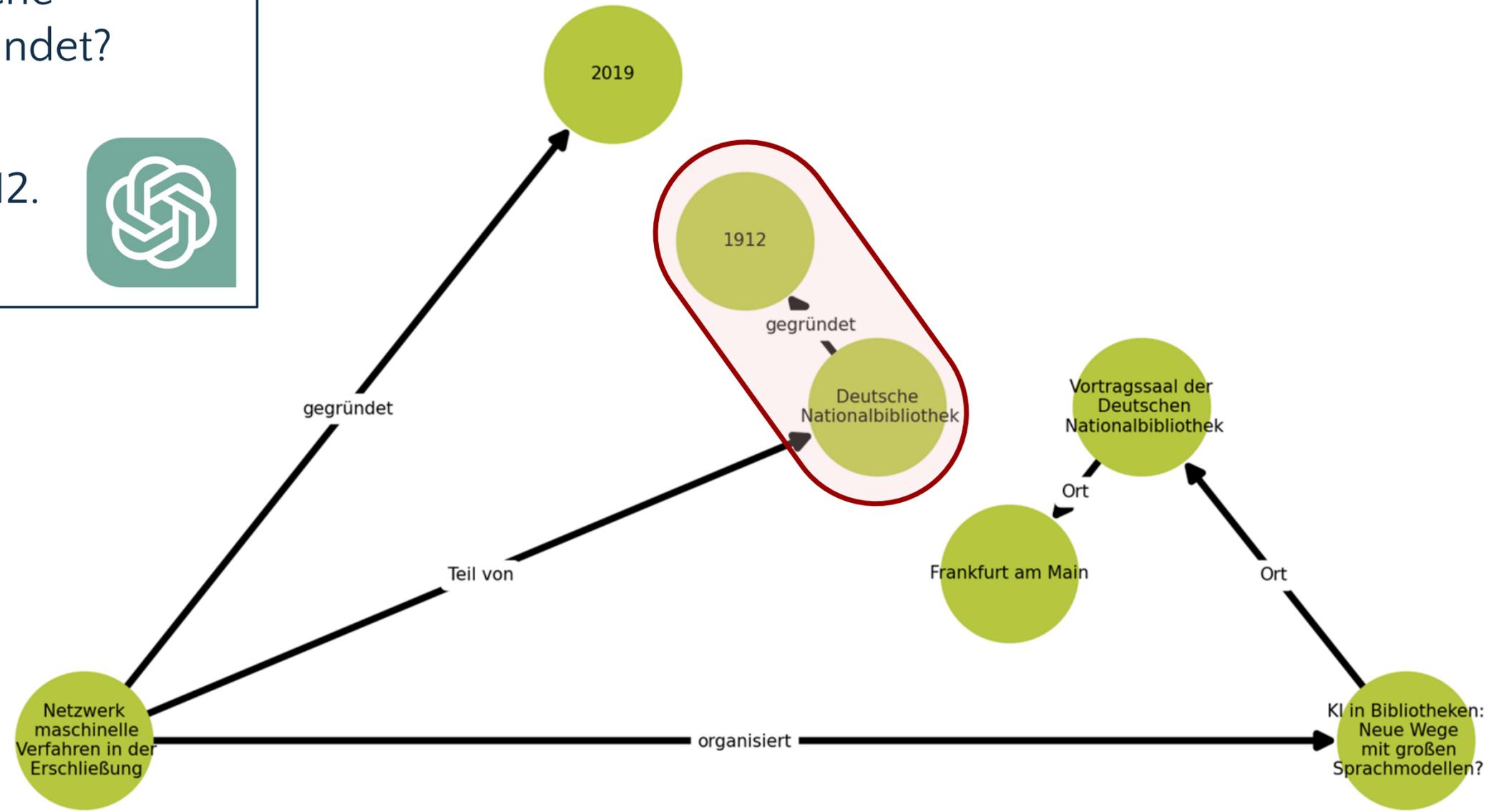
➤ Können LLMs das Wissen im Wissensgraph nutzen?

Anfrage an unseren Beispiel-Wissensgraph



Wann wurde die Deutsche Nationalbibliothek gegründet?

1912.



Suchmaschinen ↔ Wissensgraphen

Komplexe Anfragen beantworten: Bing Chat

In welchem Jahr wurde das Netzwerk, das die Veranstaltung "KI in Bibliotheken: Neue Wege mit großen Sprachmodellen?" organisiert, gegründet?

- ✓ Searching for: **KI in Bibliotheken: Neue Wege mit großen Sprachmodellen? Netzwerk Gründungsjahr**
- ✓ Generating answers for you...

Das Netzwerk, das die Veranstaltung "KI in Bibliotheken: Neue Wege mit großen Sprachmodellen?" organisiert, ist das "Netzwerk maschinelle Verfahren in der Erschließung". Es wurde gegründet, um die Anwendung von maschinellen Verfahren in Bibliotheken zu fördern und zu unterstützen ¹ ². Leider konnte ich keine Informationen darüber finden, wann genau das Netzwerk gegründet wurde.

Die Veranstaltung "KI in Bibliotheken: Neue Wege mit großen Sprachmodellen?" wird von der Deutschen Nationalbibliothek organisiert und findet am 7. und 8. Dezember 2023 in Frankfurt am Main statt ¹. Sie befasst sich mit den aktuellen Entwicklungen im Bereich generativer KI und großer Sprachmodelle und deren Bedeutung für Bibliotheken. Die Veranstaltung findet im Rahmen des "Netzwerks maschinelle Verfahren in der Erschließung" statt ¹ ².

Ich hoffe, das hilft Ihnen weiter.

Learn more: [1. bibliotheksportal.de](#) [2. wiki.dnb.de](#) [3. bsi.bund.de](#) [4. acatech.de](#) [5. wiki.dnb.de](#)

Wieso hat Bing keine Antwort?

Von welchem Netzwerk wird die Veranstaltung "KI in Bibliotheken: Neue Wege mit großen Sprachmodellen?" organisiert?

- ✓ Searching for: **KI in Bibliotheken: Neue Wege mit großen Sprachmodellen Netzwerk**
- ✓ Generating answers for you...

Die Veranstaltung "KI in Bibliotheken: Neue Wege mit großen Sprachmodellen?" wird vom **Netzwerk maschinelle Verfahren in der Erschließung** organisiert ¹. Die Fachtagung findet am **7. und 8. Dezember 2023** in der **Deutschen Nationalbibliothek in Frankfurt am Main** statt ^{1 2}.

Ich hoffe, das hilft Ihnen weiter!

Learn more: [1. wiki.dnb.de](#) [2. bibliothek](#)

The screenshot shows a Confluence page with the following content:

- Pages / Netzwerk maschinelle Verfahren in der Erschließung
- Created by Schöning-Walter, Christa, last modified on 2023-11-14
- Willkommen** zur Fachtagung 2023 des **Netzwerks maschinelle Verfahren in der Erschließung** am **07. und 08. Dezember 2023**
- Schwerpunkt der **Fachveranstaltung der Deutschen Nationalbibliothek** sind die aktuellen Entwicklungen im Bereich generativer KI und großer Sprachmodelle und deren Bedeutung für die Bibliotheken.

In welchem wurde das Netzwerk maschinelle Verfahren in der Erschließung gegründet?

- ✓ Searching for: **Netzwerk maschinelle Verfahren in der Erschließung gründungsjahr**
- ✓ Generating answers for you...

Das Netzwerk maschinelle Verfahren in der Erschließung wurde im Jahr **2019** gegründet ¹. Es ist ein Rahmen für den Informations- und Erfahrungsaustausch mit forschenden, entwickelnden und anwendenden Institutionen, die übergreifende Zusammenarbeit und den Ergebnistransfer fördern ¹. Die Deutsche Nationalbibliothek nutzt seit etwa zehn Jahren Technologien aus den Bereichen Natural Language Processing (NLP), Computerlinguistik und Machine Learning für die inhaltliche Erschließung von Medienwerken ¹. Die Grundlage ihrer maschinellen Analyseprozesse sind unstrukturierte Texte, die digital vorliegen, beispielsweise die Volltexte oder Inhaltsverzeichnisse gesammelter Publikationen ¹.

¹: Quelle

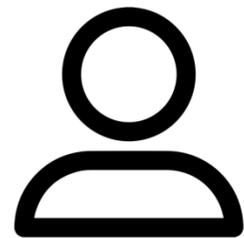
Learn more: [1. wiki.dnb.de](#) [2. bing.com](#)

The screenshot shows a Confluence page with the following content:

- Pages
- Netzwerk maschinelle Verfahren in der Erschließung**
- Created by Wagner, Nico, last modified by Schöning-Walter, Christa on 2023-11-03
- Buttons for: Fachtagung 2019, Workshop 2020, Fachtagung 2021, Workshop 2022, and **Fachtagung 2023**.
- Information box: **Fachtagung am 7. und 8. Dezember 2023: KI in Bibliotheken - Neue Wege mit großen Sprachmodellen?** Das Programm ist jetzt online.

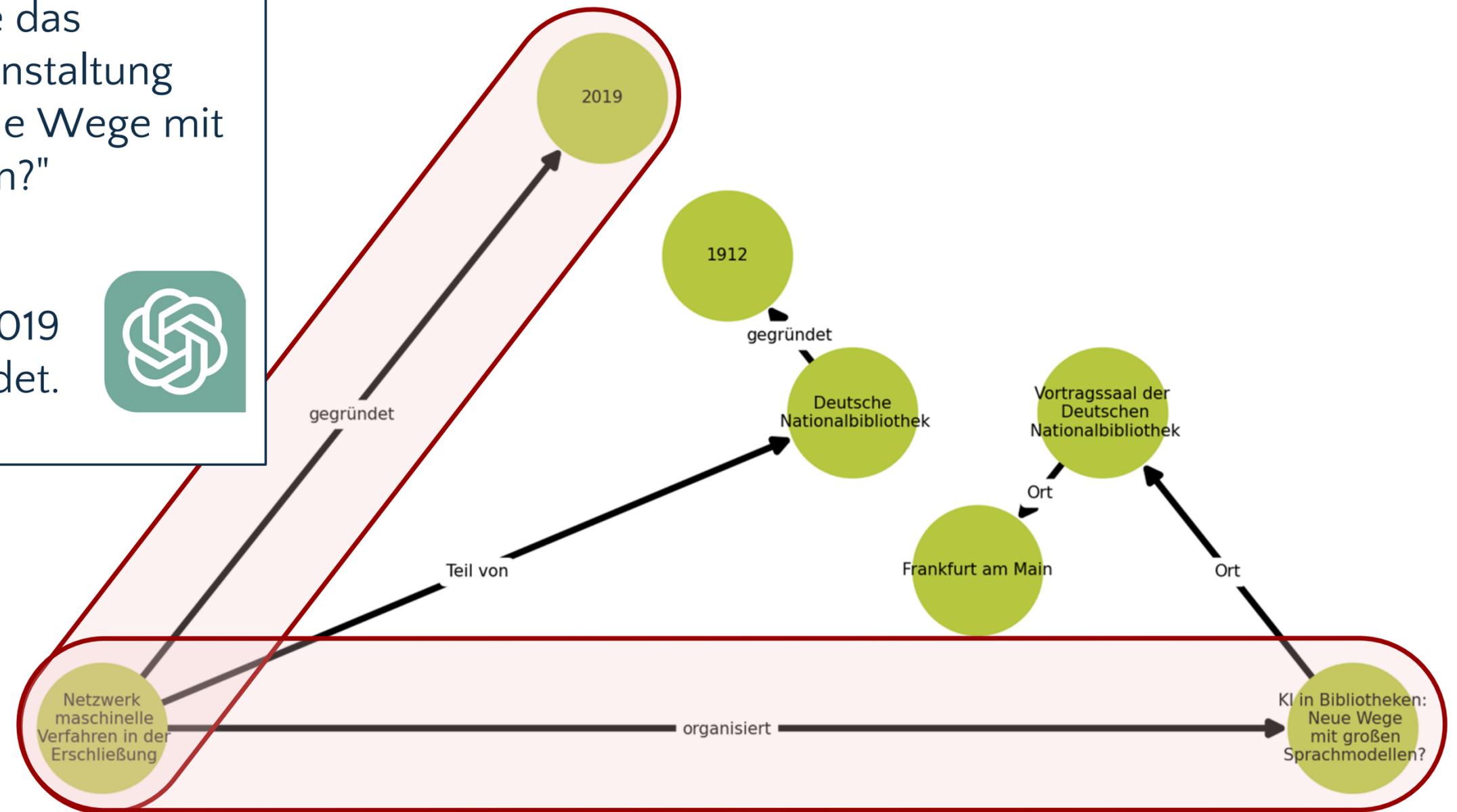
Suchmaschinen ↔ Wissensgraphen

Komplexe Anfragen beantworten: Unser Beispielwissensgraph

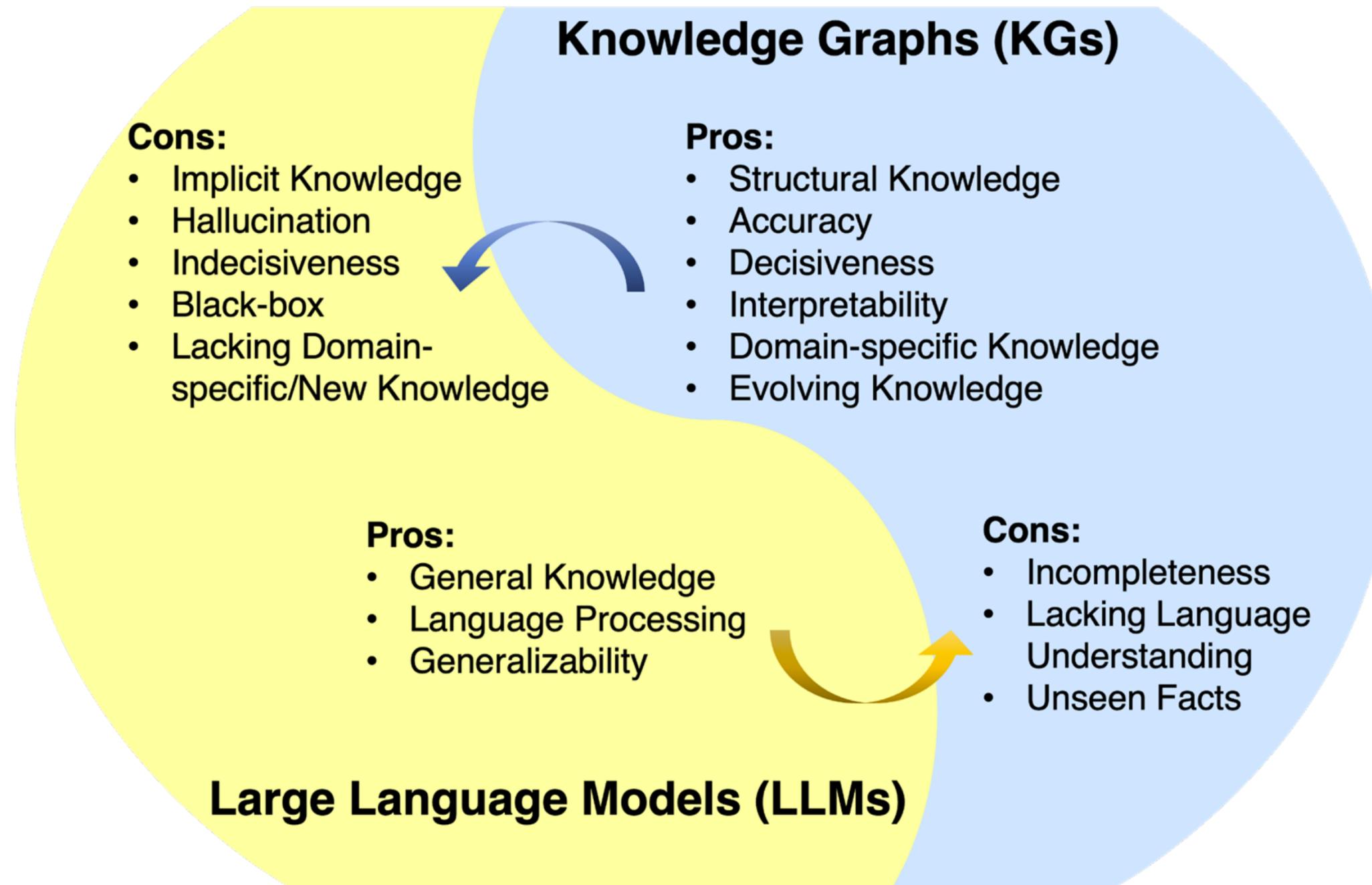


In welchem Jahr wurde das Netzwerk, das die Veranstaltung "KI in Bibliotheken: Neue Wege mit großen Sprachmodellen?" organisiert, gegründet?

Das Netzwerk wurde 2019 gegründet.



Wissensgraphen ↔ LLMs



Pan, Shirui, et al. "Unifying Large Language Models and Knowledge Graphs: A Roadmap." arXiv preprint arXiv:2306.08302 (2023).

Zwischenfazit: Gegenüberstellung Wissensgraphen, Suchmaschinen, LLMs

Three Pillars of "Machine Knowledge"			
			
Representation:	crisp	mixed	<i>latent</i>
Correctness:	high	mixed	mixed
Freshness:	mostly	fresh	stale
Coverage:	limited	everything	huge
Cost/Query:	low	low	high
Provenance:	tangible	tangible	intangible
			
Coolness:	nerdy	mundane	amazing

Keynote "Knowledge Graphs in the Age of Language Models. KG Blind Spots and What Could Be Done About Them" von Gerhard Weikum bei der International Semantic Web Conference 2023

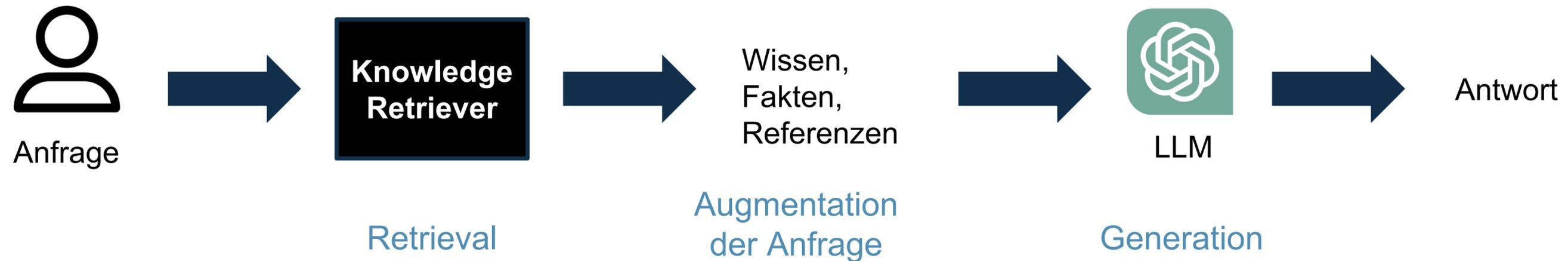
Wie kann ein LLM Dinge nachschlagen?

LLMs vs. RAGs

- Standard-LLMs:

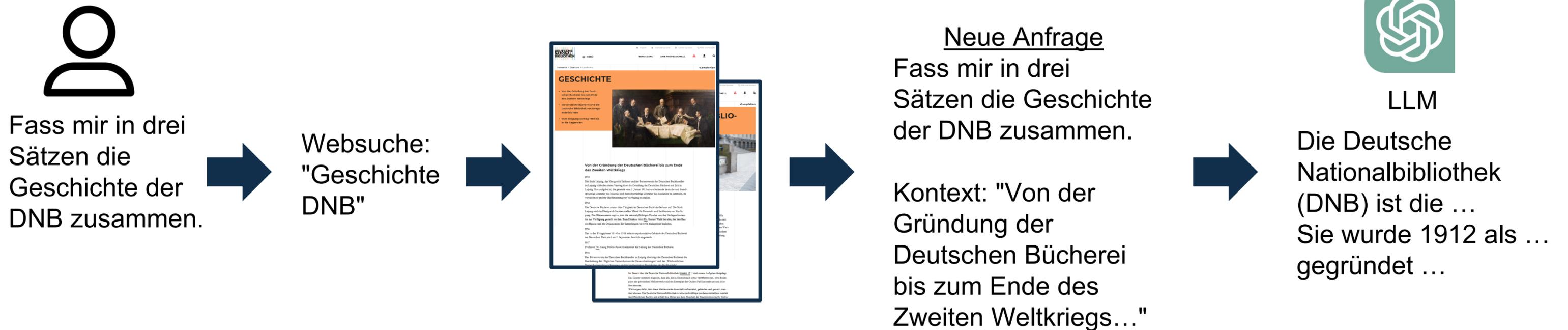


- LLMs mit Retrieval Augmented Generation (RAG):



Retrieval Augmented Generation mit Suchmaschinen

- Retrieval-basiertes LLM: Ein LLM, das zur Beantwortung einer Anfrage eine externe Quelle verwendet.
 - Retriever: Suchmaschine.
 - Zusätzlicher Kontext: Texte von Webseiten



Knowledge Retriever für Webseiten: Information Retrieval

Klassisch / Sparse Retrieval

- Klassische Methoden TF-IDF und BM25 basierend auf Worthäufigkeiten

Anfrage: "Wann wurde die DNB gegründet"



[0 0 1 0 1 ...]

Website: "Die DNB wurde 1912 gegründet."



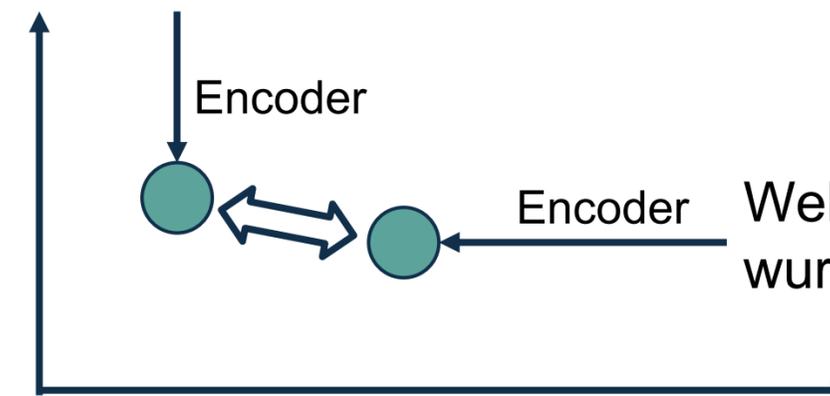
[1 0 1 0 1 ...]

➤ Lexikalische Überschneidung

Semantische Suche / Dense Retrieval

- Semantische Repräsentation von Anfrage und Website z. B. mit Word Embeddings
- Vektorprodukt als Ähnlichkeitsmaß

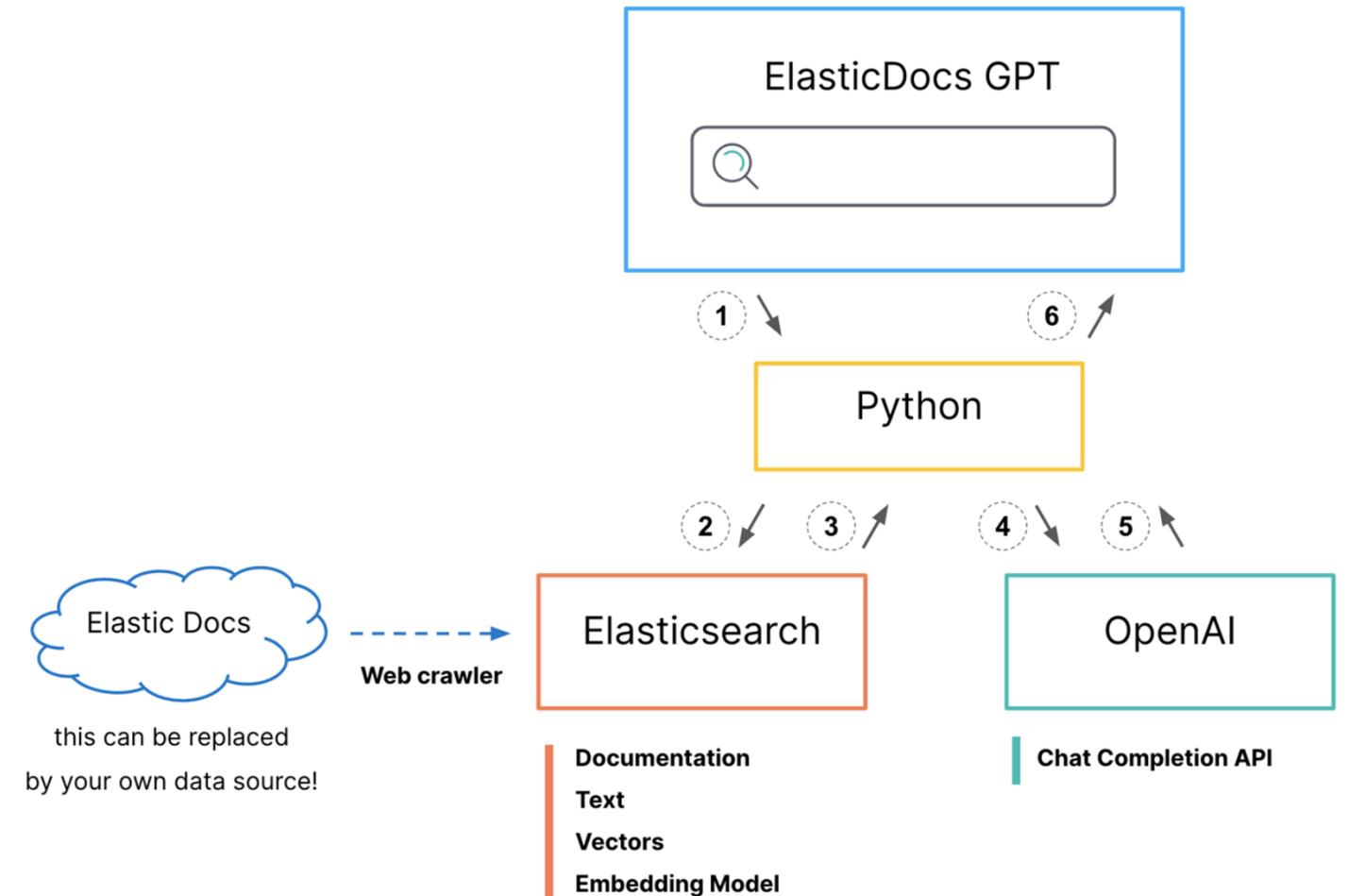
Anfrage: "Wann wurde die DNB gegründet"



Website: "Die DNB wurde 1912 gegründet."

Programmierung

- ElasticSearch: Etablierte Suchmaschine für eigene Dokumente
- Verbindung von ElasticSearch mit ChatGPT über Python-Framework



<https://www.elastic.co/search-labs/blog/articles/chatgpt-elasticsearch-openai-meets-private-data>

Beispiel: CORE-GPT

- CORE-GPT beantwortet Fragen mit Hilfe wissenschaftlicher Referenzen
- Im Gegensatz zu Galactica: Retrieval Augmented Generation
 - Keine halluzinierten Referenzen

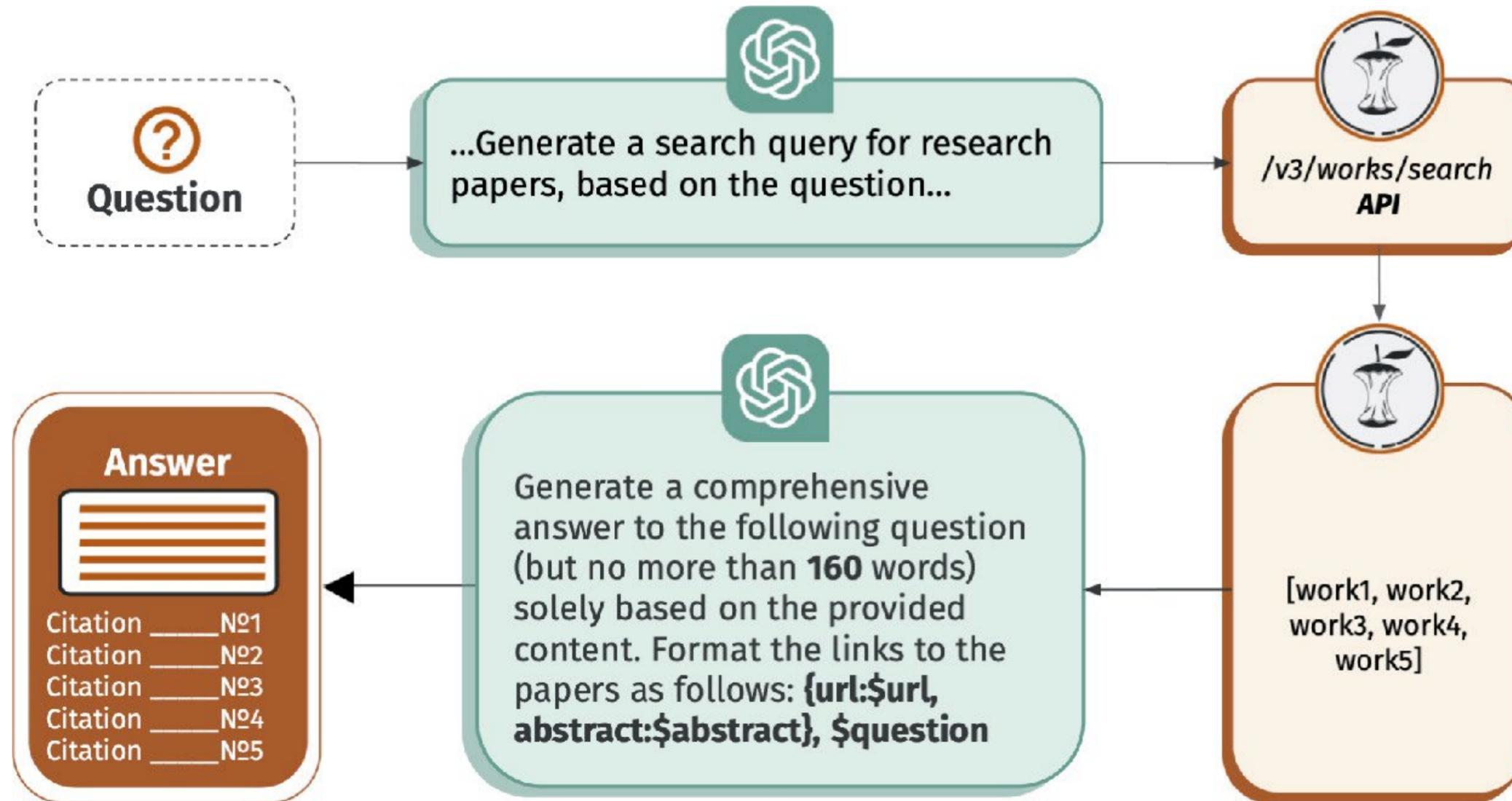
What are the latest research papers in large language models including gpt4

[Ask again](#)

GPT-4 is a large-scale, multimodal model that accepts image and text inputs, producing text outputs. It demonstrates human-level performance in various professional and academic benchmarks, such as passing a simulated bar exam with top 10% scores. The development of GPT-4 involved creating infrastructure and optimization methods that function predictably across different scales [1]. TextBox 2.0 is a comprehensive and unified library focusing on the use of pre-trained language models (PLMs) for text generation [2]. Large language models (LLMs) have shown outstanding performance in natural language processing tasks, and a recent bibliometric review of LLMs research was conducted from 2017 to 2023, providing a roadmap for current research trends [3]. Furthermore, a survey of LLMs [4] and a generative design approach utilizing generative pre-trained language models for biologically inspired design concept generation have also been published [5].

1. <https://core.ac.uk/works/142057426> - GPT-4 Technical Report
2. <https://core.ac.uk/works/137357969> - TextBox 2.0: A Text Generation Library with Pre-trained Language Models
3. <https://core.ac.uk/works/142129336> - A Bibliometric Review of Large Language Models Research from 2017 to 2023
4. <https://core.ac.uk/works/142115282> - A Survey of Large Language Models
5. <https://core.ac.uk/works/137358742> - Biologically Inspired Design Concept Generation Using Generative Pre-Trained Transformers

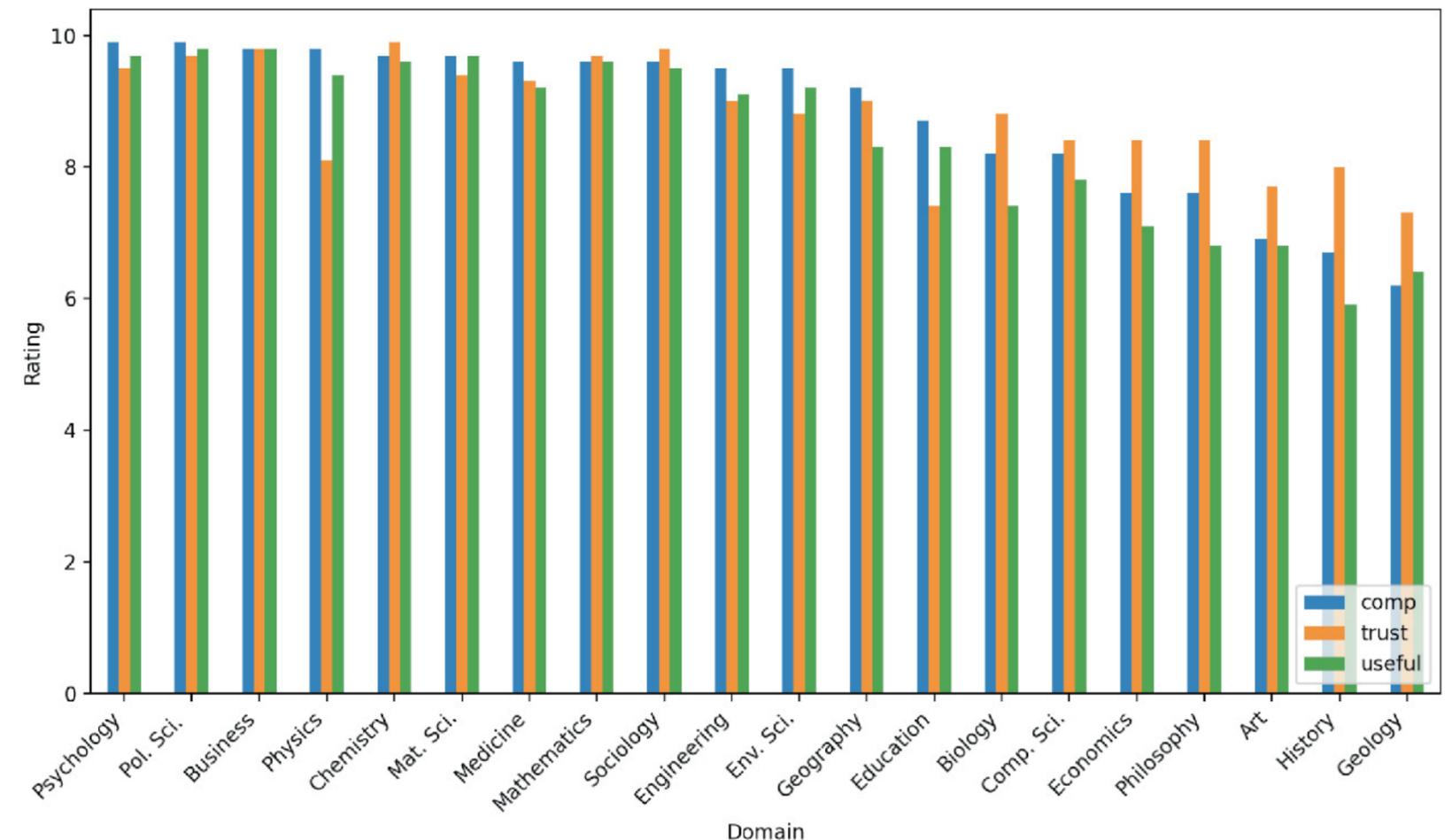
CORE-GPT: Workflow



CORE-GPT: Evaluierung

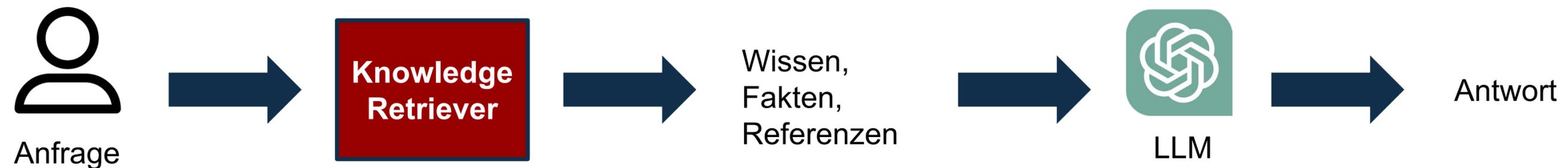
Pride, David, Matteo Cancellieri, and Petr Knoth. "CORE-GPT: Combining Open Access research and large language models for credible, trustworthy question answering." International Conference on Theory and Practice of Digital Libraries. Cham: Springer Nature Switzerland, 2023.

- In 75% der Themenfelder wurden Umfang, Vertrauen und Nutzen der Antworttexte von mindestens mit 8/10 Punkten bewertet
 - Die Antworten von CORE-GPT sind relevant, faktenbasiert und beinhalten Referenzen



Zuverlässige RAG-LLMs: Ablauf

- Welche Schritte muss ein RAG-LLM im Detail durchlaufen, um zuverlässige Ergebnisse zu generieren?
- Anforderungen:
 - Korrektheit der Antwort
 - Bereitstellung von Referenzen
 - Präzision der Antwort
- Insbesondere wichtig in kritischen Themen wie Medizin
- Verfeinerung des Knowledge Retrievers



Beispiel-Ablauf: Suchanfrage

- Welcher Teil der Nutzeranfrage ist relevant für das Retrieval?

Hallo! :) Ich fahre nächsten Donnerstag zu einer Veranstaltung der Deutschen Nationalbibliothek. Bei der Veranstaltung geht es um die Rolle von Large Language Models in Bibliotheken. Spannend! Aber eigentlich wollte ich nur wissen, in welcher Stadt die DNB ist. Kannst du mir das sagen?

✓ Searching for: **Deutsche Nationalbibliothek Stadt**

- Anfrageoptimierung: Was für Suchbegriffe verwenden?

Wo befindet sich die Deutsche Nationalbibliothek?

✓ Searching for: **Deutsche Nationalbibliothek location**

- Personalisierung der Anfrage

Wie komme ich am besten nach Frankfurt?

✓ Searching for: **best way to travel to Frankfurt from Hannover**

Beispiel-Ablauf: Auswertung der Ergebnisse

- Wie kann das RAG-LLM sicherstellen, dass gefundene Ergebnisse relevant sind?
- Beispiel: Wissenschaftliche Referenzen
 - Sind die Publikationen in vertrauenswürdigen Konferenzen/Journalen veröffentlicht?
 - Sind die Publikationen aktuell?
 - Gibt es Konflikte in den Referenzen?
- Sind die gefundenen Informationen ausreichend zum Beantworten der Anfrage? Wenn nicht: LLM muss das erkennen und entsprechend reagieren
 - "Ich weiß es nicht."
 - Weiter suchen
 - Experten fragen

Gib mir aktuelle Surveys über Large Language Models.

- Liu, Yiheng, et al. "Summary of ChatGPT-related Research and Perspective Towards the Future of Large Language Models." *Meta-Radiology* (2023): 100017.

Beispiel-Ablauf: Sortierung der Ergebnisse

- Welche Ergebnisse sind am wichtigsten für die Antwort?
- Ranking von Websites: Klassische Problemstellung im Web
- Welche Referenz wird für welchen Teil der Antwort verwendet?

Kann Alkohol in Maßen förderlich für die Gesundheit sein?

✓ Searching for: **Alkohol in Maßen förderlich für die Gesundheit**

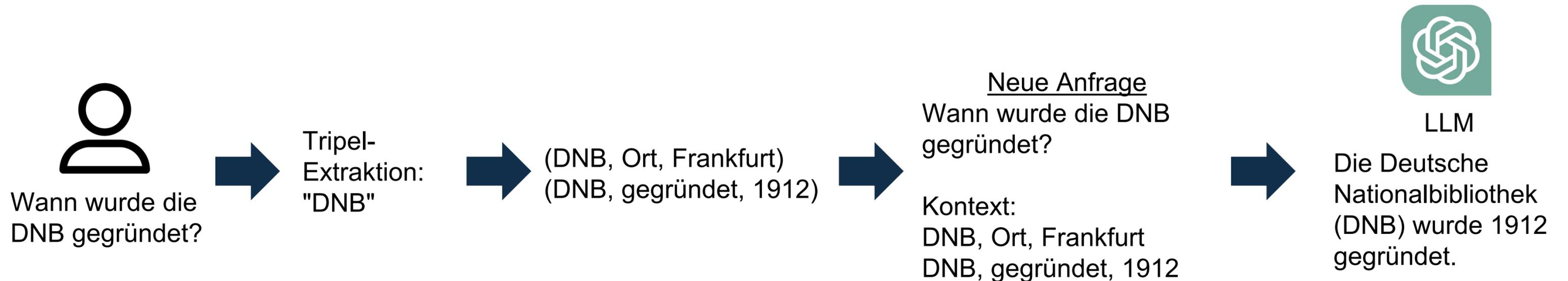
✓ Generating answers for you...

Ja, es gibt Studien, die darauf hinweisen, dass ein moderater Alkoholkonsum das Risiko für Herz-Kreislauf-Erkrankungen senken kann ^{1 2 3}. Allerdings gibt es auch Studien, die diese Aussage widerlegen ⁴. Es ist wichtig zu beachten, dass Alkohol in höheren Mengen schädlich sein kann und dass die gesundheitlichen Vorteile von moderatem Alkoholkonsum von vielen Faktoren abhängen, wie z.B. Alter, Geschlecht, Gewicht und Gesundheitszustand ^{1 2 3 4 5}. Es ist daher ratsam, sich an die Empfehlungen der Weltgesundheitsorganisation (WHO) zu halten, die einen maximalen Alkoholkonsum von 20 Gramm reinem Alkohol pro Tag für Frauen und 30 Gramm für Männer empfiehlt ⁵. Bitte beachten Sie, dass Alkoholkonsum gesundheitliche Risiken birgt und dass Sie sich bei Fragen zu Ihrer Gesundheit immer an einen Arzt oder eine Ärztin wenden sollten.

Learn more: [1. aerztezeitung.de](#) [2. bing.com](#) [3. gesundheitsjournal.de](#) [4. gala.de](#) [+1 more](#)

Retrieval Augmented Generation mit Wissensgraphen

- Retriever: Triple-Extraktion / Teilgraph-Suche
- Zusätzlicher Kontext: Tripel



Knowledge Retriever für Wissensgraphen

- Wissensgraphen enthalten Millionen von Entitäten und Relationen
- Problem: Größe von Wissensgraphen übersteigt die Anzahl möglicher Token im Anfrageprompt
 - Wie kann der für die Anfrage relevante Teil eines Wissensgraphen extrahiert werden?

1. Identifikation von Entitäten in der Anfrage
2. Extraktion von Tripeln im Wissensgraph, die mit diesen Entitäten verbunden sind



Ausblick: LLMs in Bibliotheken

Fallbeispiel: Automatisierter Bibliothekar

- Der Katalog der DNB enthält Metadaten zu 46,3 Millionen Medienwerken
- Können LLMs helfen, einen einfacheren Zugriff auf die Medienwerke im DNB-Katalog zu erhalten?
- LLMs als automatisierter Bibliothekar?

The screenshot shows the search interface of the Deutsche Nationalbibliothek (DNB) catalog. At the top left is the logo 'DEUTSCHE NATIONAL BIBLIOTHEK'. To the right are navigation links: Kontakt, A-Z, Träger / Förderer, Datenschutz, Impressum, Hilfe, Mein Konto, and English. Below the logo is a sidebar menu with options: Katalog (selected), Einfache Suche, Erweiterte Suche, Browsen (DDC), Suchverlauf, Meine Auswahl, Hilfe, Datenschop, Mein Konto, Ablieferung von Netzpublikationen, and Informationsvermittlung. A 'Login' button is also present. The main search area is titled 'KATALOG DER DEUTSCHEN NATIONALBIBLIOTHEK' and features tabs for 'Gesamter Bestand', 'Musikarchiv', 'Exilsammlungen', and 'Buchmuseum'. The search form includes fields for 'Titel', 'Person', 'Inhaltsverzeichnis', and 'Schlagwörter', each with a dropdown menu and a search icon. There are also 'und' dropdowns for combining search terms. A 'Jahr' field is present with a range example: 'zum Beispiel: 1948-1980 oder 1976- oder 1955'. Below this is a section for 'Neu verzeichnet in der Deutschen Nationalbibliografie in den Reihen:' with checkboxes for A, B, C, G, H, M, T, O, N, and options for 'Alle' and 'Keine'. A date range field is also visible, with '10.11.2023' entered. A 'Finden' button is at the bottom of the search form. At the bottom of the page, there are tabs for 'Materialarten', 'Standorte/Kataloge', and 'Sachgruppen', with 'Alle' and 'Keine' options.

Suchportal des Katalogs der DNB

Automatisierter Bibliothekar: Beispiel

- Das Forschungszentrum L3S veröffentlicht seit 2018 regelmäßig die Zeitschrift "Binaire" zu verschiedenen Themen der Informatik
- Wie kann ein LLM und der Katalog der DNB helfen, Informationen über die verschiedenen Binaire-Ausgaben zu erhalten?



Automatisierter Bibliothekar: Beispiel - Datenextraktion

- Der DNB-Katalog unterstützt den Export von Suchergebnissen als Wissensgraph

Ergebnis der Suche nach: "l3s" and "binaire" im Bestand: Gesamter Bestand

1 - 10 von 16 Datum (neuestes zuerst) ▼ sortieren →

- 1 Binaire / 2023, No. 1 Personalisierte Medizin 2023
- 2 Binaire / 2023, No. 2 Generative künstliche Intelligenz 2023
- 3 Binaire / 2022/1 L3S und Partner 2022
- 4 Binaire / 2021/1 Sonderausgabe - Startup 2021
- 5 Binaire / 2021/2 Digitales Lernen 2021

Suche

Ausgewählte Datensätze

Suchanfrage	Einschränkungen
<input checked="" type="radio"/> "l3s" and "binaire"	- keine

Datenformate

Katalog	Format	Zeichensatz	Anmerkung
<input type="radio"/>	JsonLd	UTF-8	
<input type="radio"/>	MARC21	UTF-8	
<input type="radio"/>	MARC21-xml	UTF-8	
<input type="radio"/>	RDFxml	UTF-8	
<input checked="" type="radio"/>	Turtle	UTF-8	
<input type="radio"/>	csv_title	UTF-8	Das Format ist nur für Titeldaten verfügbar
<input type="radio"/>	oai_dc	UTF-8	Das Format ist nur für Titeldaten verfügbar
<input type="radio"/>	pdf-natbib	UTF-8	Das Format ist nur für Titeldaten verfügbar

Auftrag starten

Datenexport

```

dcterms:license
<http://creativecommons.org/publicdomain/zero/1.0/>;
  dcterms:modified
  "2023-05-30T07:16:17.000"^^xsd:dateTime
  .
<https://d-nb.info/1281779857>
dcterms:issued "2023";
  rdau:P60493 "2023, No. 1
  Personalisierte Medizin";
  dcterms:isPartOf
  <https://ld.zdb-services.de/resource/3028276-7>;
  owl:sameAs
  <http://hub.culturegraph.org/resource/DNB-1281779857> .
<https://d-nb.info/1305986296> a
  bibo:Document;
  dcterms:medium
  <http://rdaregistry.info/termList/RDA/CarrierType/1044>;
  rdau:P60493
  
```

Exportierter Wissensgraph

Automatisierter Bibliothekar: Beispiel - Wissensgraph

- Der DNB-Katalog unterstützt den Export von Suchergebnissen als Wissensgraph

```
index_creator = GraphIndexCreator(llm=OpenAI(temperature=0))
graph = index_creator.from_text('')

triple_strings = set()
for s, p, o in g:
    if p not in property_label_dict:
        continue
    property_label = property_label_dict[p]

    if s in node_label_dict:
        s_label = node_label_dict[s]
    else:
        s_label = str(s)

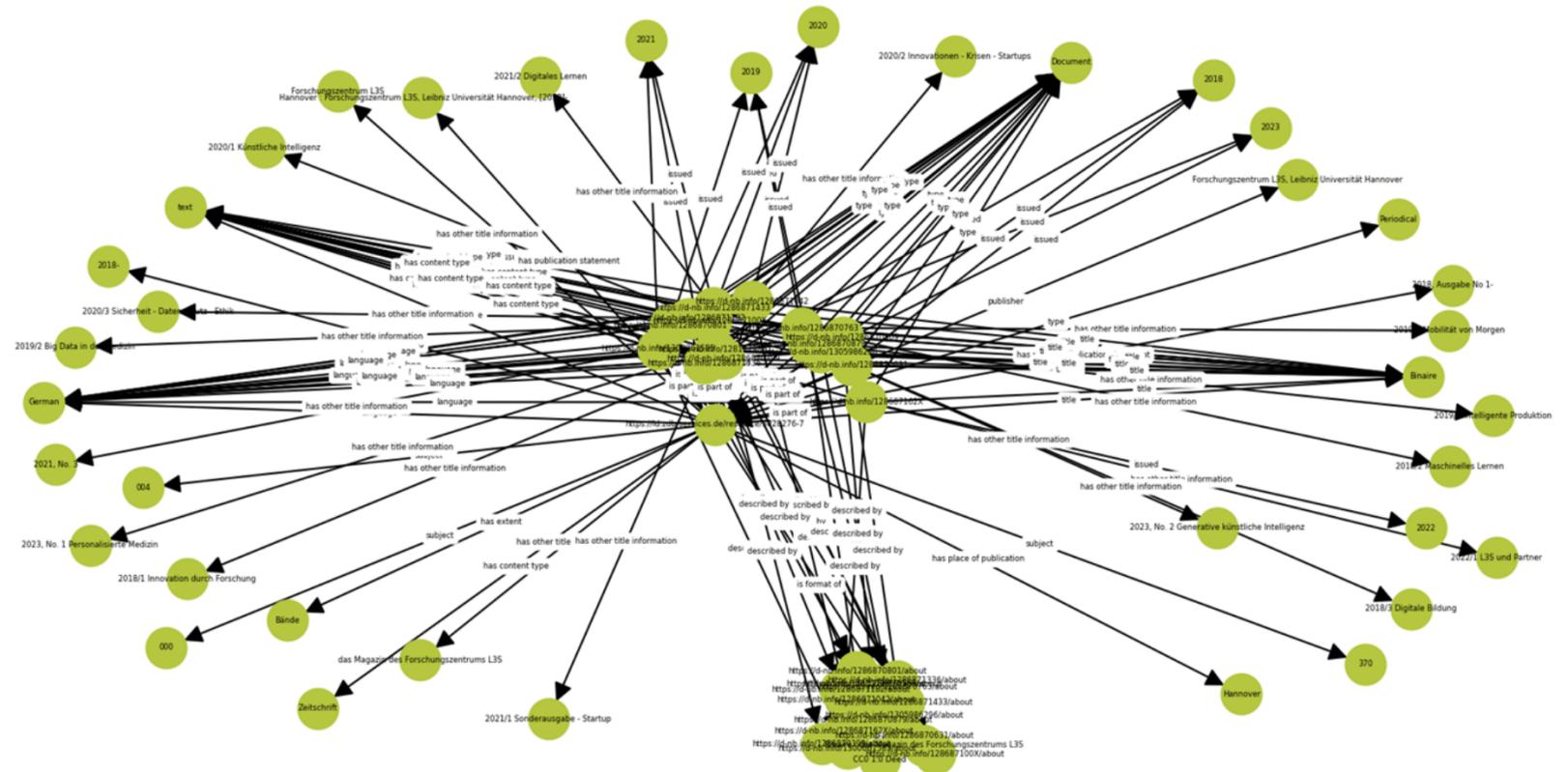
    if o.__class__ == Literal:
        o_label = str(o)
    else:
        if o in node_label_dict:
            o_label = node_label_dict[o]
        else:
            o_label = str(o)

    graph.add_triple(KnowledgeTriple(s_label, property_label, o_label))

...

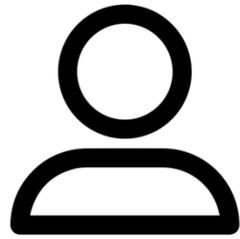
question = 'Nenn mir drei beispielhafte Titel von Binaire-Ausgaben.'
chain = GraphQChainWithoutEntityExtraction.from_llm(OpenAI(temperature=0),
                                                    graph=graph, verbose=True, qa_prompt=GRAPH_QA_PROMPT_DE)
res = chain.run(question)
print(res)
```

Programmierung mit Python/LangChain



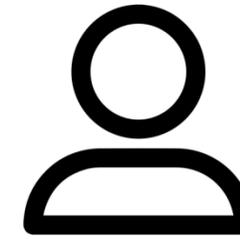
Visualisierung des Wissensgraphs

Automatisierter Bibliothekar: Beispiel - Anfragen



Nenn mir drei beispielhafte Titel von Binaire-Ausgaben.

2023, No. 1 Personalisierte Medizin;
2020/1 Künstliche Intelligenz; 2021/1
Sonderausgabe - Startup.



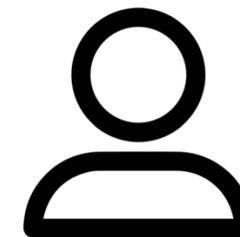
Wer ist der Herausgeber von Binaire?

Der Herausgeber von Binaire ist das
Forschungszentrum L3S.



Welche Ausgaben von Binaire behandeln das Thema Medizin?

Die Ausgaben von Binaire, die das
Thema Medizin behandeln, sind
2023, No. 1 Personalisierte Medizin,
2019/2 Big Data in der Medizin und
2018/1 Innovation durch Forschung.



Unter welcher Lizenz wird die Binaire veröffentlicht?

Alle Ausgaben der Binaire werden
unter der Lizenz CC0 1.0 Deed
veröffentlicht.



Strukturierung von Bibliotheksbeständen

- Können LLMs helfen,
 - strukturierte Metadaten über Medienwerke zu erhalten?
 - Zusammenfassungen über Medienwerke zu generieren?
 - Suchanfragen zu generieren?

GO You
Generiere RDF-Tripel, die Metadaten zum Paper "Attention is all you need" beschreiben. Nutze dazu das dcterms-Vokabular.

```
turtle
@prefix dcterms: <http://purl.org/dc/terms/>.
@prefix ex: <http://example.org/>.

ex:Paper1
  a dcterms:BibliographicResource;
  dcterms:title "Attention is All You Need";
  dcterms:creator "Vaswani, Ashish" ;
  dcterms:contributor "Shazeer, Noam";
  dcterms:contributor "Parmar, Niki";
  dcterms:contributor "Uszkoreit, Jakob";
  dcterms:contributor "Jones, Llion";
  dcterms:contributor "Gomez, Aidan N.";
  dcterms:contributor "Kaiser, Lukasz";
  dcterms:contributor "Polosukhin, Illia";
  dcterms:date "2017-06-12";
  dcterms:publisher "arXiv";
  dcterms:subject "Attention mechanism";
```

GO You
Schreibe eine SPARQL-Anfrage mit dem dcterms-Vokabular, um Veröffentlichungen über das Thema "Medizin" zu finden.

```
sparql

PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT ?publication ?title ?creator ?date
WHERE {
  ?publication a dcterms:BibliographicResource ;
               dcterms:subject "Medicine" ;
               dcterms:title ?title ;
               dcterms:creator ?creator ;
               dcterms:date ?date .
}

ORDER BY ?date
```

Offene Fragen und Herausforderungen

Offene und subjektive Anfragen

- Wo kann RAG nicht helfen?
 - Einbindung von RAG führt zu keiner signifikanten Verbesserung von LLMs bei offenen Fragen

Model	MAUVE↑	PPL _{GPT-3} ↓	RankGen↑	EntityF1↑	SeqRep_1↓
<i>kNN-LM with and without retrieval from Alon et al. (2022)</i>					
GPT-2 small (no retrieval)	0.773	13.1	11.7	0.14	0.57
GPT-2 small (+ retrieval)	0.793	14.8	11.7	0.13	0.53

Warum gefällt mir das Buch "Die Einladung" nicht?

✓ Searching for: Warum gefällt mir das Buch 'Die Einladung' nicht?
✓ Generating answers for you...

Es tut mir leid, dass dir das Buch "Die Einladung" nicht gefällt. Ich kann dir jedoch nicht sagen, warum es dir nicht gefällt, da ich keine Informationen darüber habe, was du an dem Buch nicht magst. Wenn du möchtest, kann ich dir jedoch einige allgemeine Gründe nennen, warum ein Buch einem Leser nicht gefallen könnte. Hier sind einige Gründe, die ich gefunden habe:

- Der Schreibstil des Autors könnte nicht ansprechend sein.
- Die Handlung könnte langweilig oder vorhersehbar sein.
- Die Charaktere könnten flach oder uninteressant sein.
- Das Buch könnte zu lang oder zu kurz sein.
- Das Buch könnte nicht das Genre sein, das du normalerweise liest.

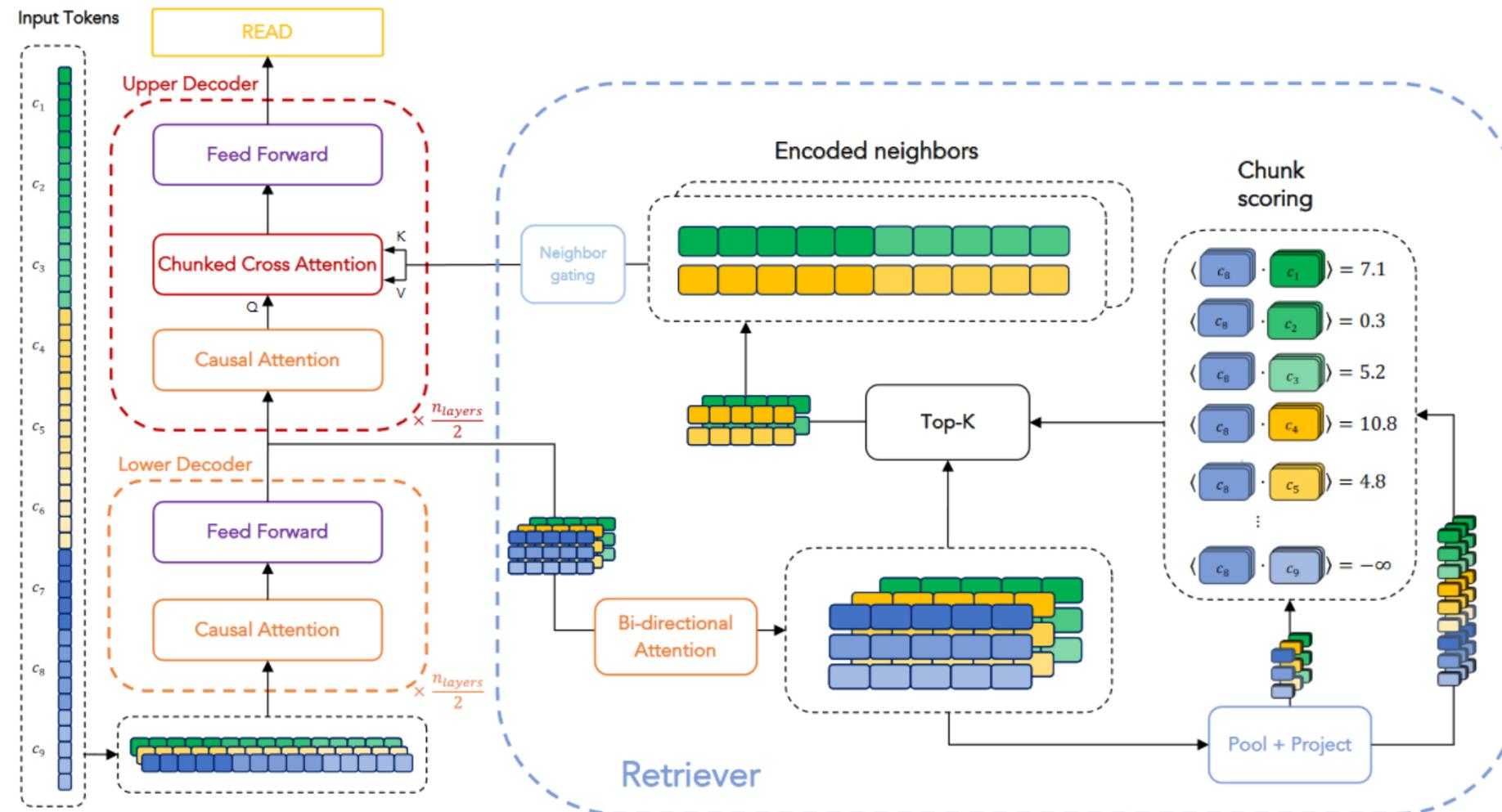
Ich hoffe, dass diese Gründe dir helfen, zu verstehen, warum dir das Buch "Die Einladung" nicht gefällt. Wenn du weitere Fragen hast, stehe ich dir gerne zur Verfügung.

- Nachfrage des Bibliothekars: "Welche Art von Büchern interessiert dich?"

Wang, Shufan, et al. "KNN-LM Does Not Improve Open-ended Text Generation." arXiv preprint arXiv:2305.14625 (2023).

Gemeinsames Training von LLMs und Retriever

- Gemeinsames Training: *Retrieval-Pretrained Transformer*
 - Gemeinsames Training eines RAG-LM für die Modellierung langer Texte

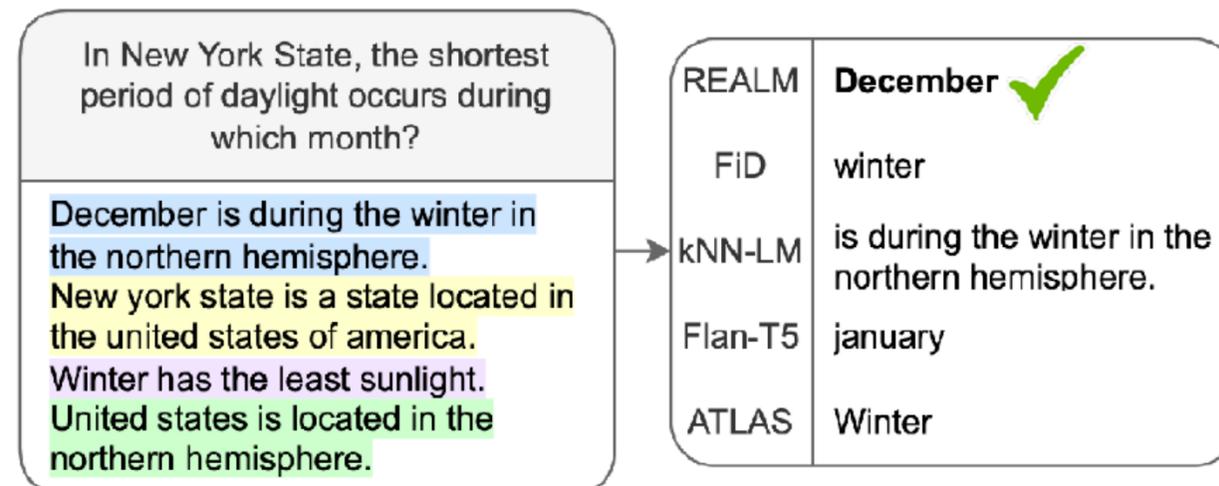


Retrieval-Pretrained Transformer: Transformer und Retriever werden gemeinsam trainiert

Rubin, Ohad, and Jonathan Berant. "Long-range Language Modeling with Self-retrieval." arXiv preprint arXiv:2306.13421 (2023).

Größe des LLMs/Datensatzes und komplexe Anfragen

- Größe des LLMs ↔ Größe des Datensatzes
 - Kleines LLM und großer Retrieval-Datensatz/Wissensgraph vs. großes LLM und kleiner Retrieval-Datensatz/Wissensgraph
 - Was sind die Vor- und Nachteile (z. B. Effizienz von Pre-training vs. Retrieval)?
- Komplexe Anfragen
 - Logische Schlüsse aus verschiedenen Anfragen und Webseiten zu ziehen ist schwierig
 - Wissensgraphen als Lösung? Hybride Ansätze?



Komplexe Anfrage und Antwort von fünf LMs

Asai, Akari et al. ACL 2023 Tutorial: Retrieval-based Language Models and Applications (<https://acl2023-retrieval-lm.github.io/>)

Fazit

1. Warum muss ein LLM Dinge nachschlagen?
 - Halluzinationen, veraltete Informationen, fehlende Referenzen, ...
 2. Wo kann ein LLM Dinge nachschlagen?
 - Internet und Wissensgraphen
 3. Wie kann ein LLM Dinge nachschlagen?
 - Retrieval-augmented generation
- Ausblick: LLMs in Bibliotheken
- Automatisierter Bibliothekar
 - Strukturierung von Bibliotheksbeständen

Kontakt



Wolfgang Nejdl



nejdl@L3S.de



www.l3s.de

 @l3s_luh

 @l3s-research-center

 @forschungszentruml3s

 @l3s-research-center

Referenzen

- Asai, Akari et al. ACL 2023 Tutorial: Retrieval-based Language Models and Applications
 - <https://acl2023-retrieval-lm.github.io/>
- Labonne, Maxime. "Improve ChatGPT with Knowledge Graphs"
 - https://mlabonne.github.io/blog/posts/Article_Improve_ChatGPT_with_Knowledge_Graphs.html?s=08
- Pan, Shirui, et al. "Unifying Large Language Models and Knowledge Graphs: A Roadmap." arXiv preprint arXiv:2306.08302 (2023).
 - <https://github.com/RManLuo/Awesome-LLM-KG>
- Pride, David, Matteo Cancellieri, and Petr Knoth. "CORE-GPT: Combining Open Access research and large language models for credible, trustworthy question answering." International Conference on Theory and Practice of Digital Libraries. Cham: Springer Nature Switzerland, 2023.
- Rubin, Ohad, and Jonathan Berant. "Long-range Language Modeling with Self-retrieval." arXiv preprint arXiv:2306.13421 (2023).
- Taylor, Ross, et al. "Galactica: A large language model for science." arXiv preprint arXiv:2211.09085 (2022).
- Wachsmuth, Henning. "NLP Research in the Age of Large Language Models"
- Wang, Shufan, et al. "KNN-LM Does Not Improve Open-ended Text Generation." arXiv preprint arXiv:2305.14625 (2023).
- ChatGPT and Elasticsearch: OpenAI meets private data
 - <https://www.elastic.co/search-labs/blog/articles/chatgpt-elasticsearch-openai-meets-private-data>
- Chatten mit Bing: Der leichte Einstieg in die Welt der künstlichen Intelligenz
 - <https://news.microsoft.com/de-de/chatten-mit-bing-der-leichte-einstieg-in-die-welt-der-kuenstlichen-intelligenz/>
- Knowledge Graphs & LLMs: Multi-Hop Question Answering
 - <https://neo4j.com/developer-blog/knowledge-graphs-llms-multi-hop-question-answering/>

Weitere Referenzen: LLMs ↔ Suchmaschinen

- Lin, Xi Victoria, et al. "RA-DIT: Retrieval-Augmented Dual Instruction Tuning." arXiv preprint arXiv:2310.01352 (2023).
- Ram, Ori, et al. "In-context retrieval-augmented language models." arXiv preprint arXiv:2302.00083 (2023).
- Shi, Weijia, et al. "REPLUG: Retrieval-augmented black-box language models." arXiv preprint arXiv:2301.12652 (2023).

Weitere Referenzen: LLMs ↔ Wissensgraphen

- Baek, Jinheon, Alham Fikri Aji, and Amir Saffari. "Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering." arXiv preprint arXiv:2306.04136 (2023).
- Hu, Linmei, et al. "A survey of knowledge enhanced pre-trained language models." IEEE Transactions on Knowledge and Data Engineering (2023).
- Pan, Jeff Z., et al. "Large language models and knowledge graphs: Opportunities and challenges." arXiv preprint arXiv:2308.06374 (2023).
- Denny Vrandečić. KGC23 Keynote: The Future of Knowledge Graphs in a World of LLMs
 - <https://www.youtube.com/watch?v=ww99npDh4cg>