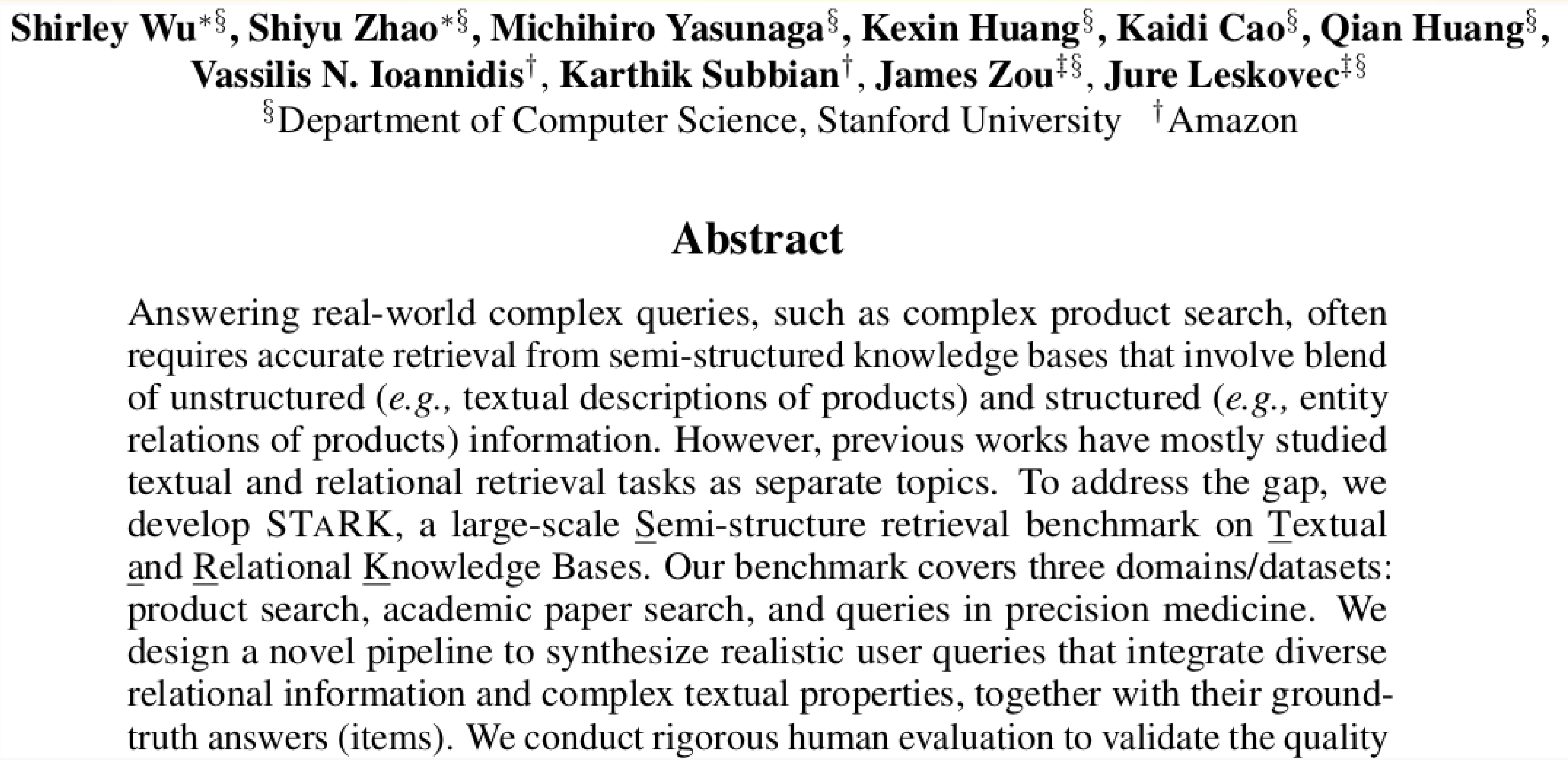
STaRK: Benchmarking LLM Retrieval on Textual and Relational Knowledge Bases

Разбор статьи
https://arxiv.org/abs/2404.13207v2

Дмитрий Колодезев @promsoft kolodezev.ru

2024.07.16 @ DataTalk

STaRK: Benchmarking LLM Retrieval



Shirley Wu[*§], **Shiyu Zhao**[*§], **Michihiro Yasunaga**[§], **Kexin Huang**[§], **Kaidi Cao**[§], **Qian Huang**[§], **Vassilis N. Ioannidis**[†], **Karthik Subbian**[†], **James Zou**[‡§], **Jure Leskovec**[‡§]

[§]Department of Computer Science, Stanford University [†]Amazon

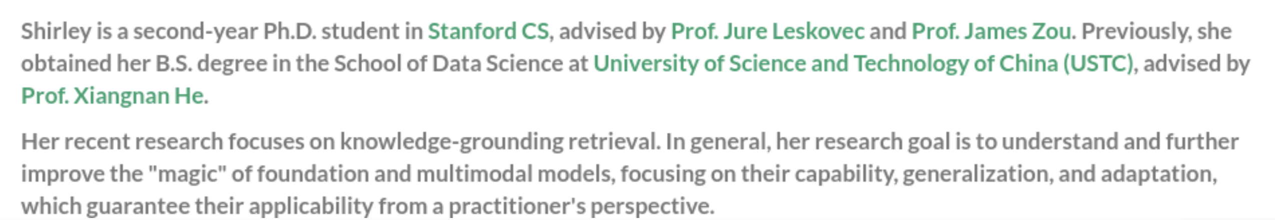
Abstract

Answering real-world complex queries, such as complex product search, often requires accurate retrieval from semi-structured knowledge bases that involve blend of unstructured (*e.g.*, textual descriptions of products) and structured (*e.g.*, entity relations of products) information. However, previous works have mostly studied textual and relational retrieval tasks as separate topics. To address the gap, we develop STaRK, a large-scale Semi-structure retrieval benchmark on Textual and Relational Knowledge Bases. Our benchmark covers three domains/datasets: product search, academic paper search, and queries in precision medicine. We design a novel pipeline to synthesize realistic user queries that integrate diverse relational information and complex textual properties, together with their ground-truth answers (items). We conduct rigorous human evaluation to validate the quality

Умницы (остальные тоже)

Shirley Wu

GraphMETRO: Mitigating Complex Graph Distribution Shifts via Mixture of Aligned Experts
Discover and Cure: Concept-aware Mitigation of Spurious Correlation (DISC)
D4explainer: in-distribution explanations of graph neural network via discrete denoising diffusion

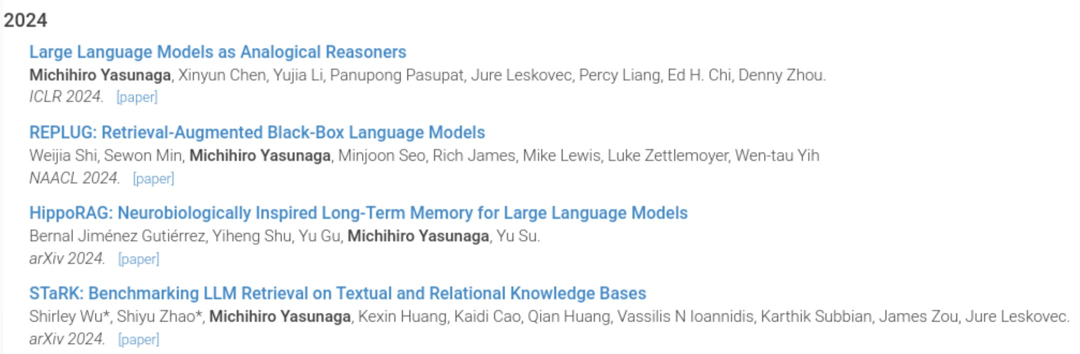


Shirley is a second-year Ph.D. student in Stanford CS, advised by Prof. Jure Leskovec and Prof. James Zou. Previously, she obtained her B.S. degree in the School of Data Science at University of Science and Technology of China (USTC), advised by Prof. Xiangnan He.

Her recent research focuses on knowledge-grounding retrieval. In general, her research goal is to understand and further improve the "magic" of foundation and multimodal models, focusing on their capability, generalization, and adaptation, which guarantee their applicability from a practitioner's perspective.

Shiyu Zhao

Michi Yasunaga



2024

Large Language Models as Analogical Reasoners
Michihiro Yasunaga, Xinyun Chen, Yujia Li, Panupong Pasupat, Jure Leskovec, Percy Liang, Ed H. Chi, Denny Zhou.
ICLR 2024. [paper]

REPLUG: Retrieval-Augmented Black-Box Language Models
Weijia Shi, Sewon Min, **Michihiro Yasunaga**, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, Wen-tau Yih
NAACL 2024. [paper]

HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models
Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, **Michihiro Yasunaga**, Yu Su.
arXiv 2024. [paper]

STaRK: Benchmarking LLM Retrieval on Textual and Relational Knowledge Bases
Shirley Wu*, Shiyu Zhao*, **Michihiro Yasunaga**, Kexin Huang, Kaidi Cao, Qian Huang, Vassilis N Ioannidis, Karthik Subbian, James Zou, Jure Leskovec.
arXiv 2024. [paper]

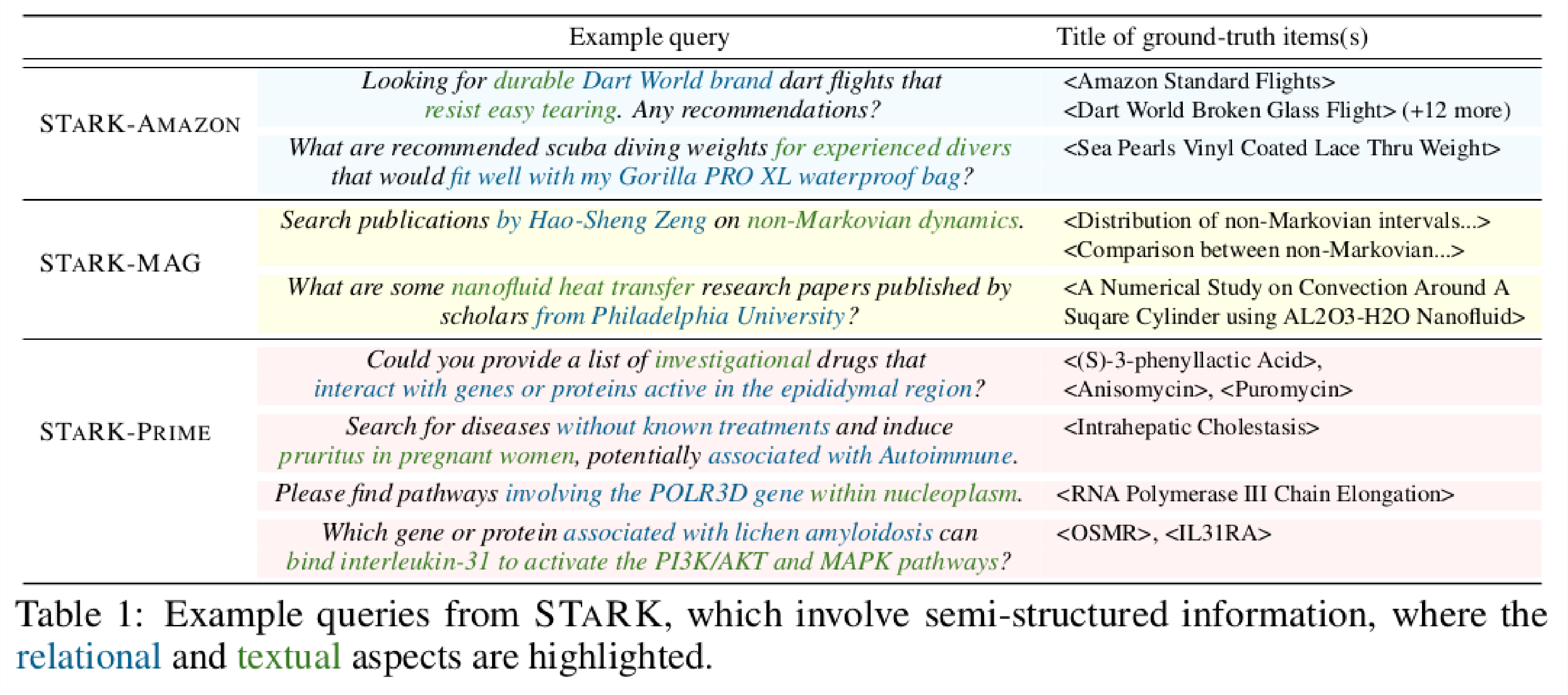
Semi-structure retrieval benchmark on Textual and Relational Knowledge Bases

- Поиск (RAG, вопросно-ответные системы) везде
- Данные организаций — тексты со связями
- Как измерить качество поиска
- Хорошо ли текущие системы + LLM справляются?

With STaRK, we found that current LLM retrieval systems CANNOT accurately retrieve information. More powerful systems are needed!

https://www.linkedin.com/posts/leskovec_reduce-llm-hallucinations-with-rag-over-textual-activity-7190745116339302401-da4n

Текст и граф



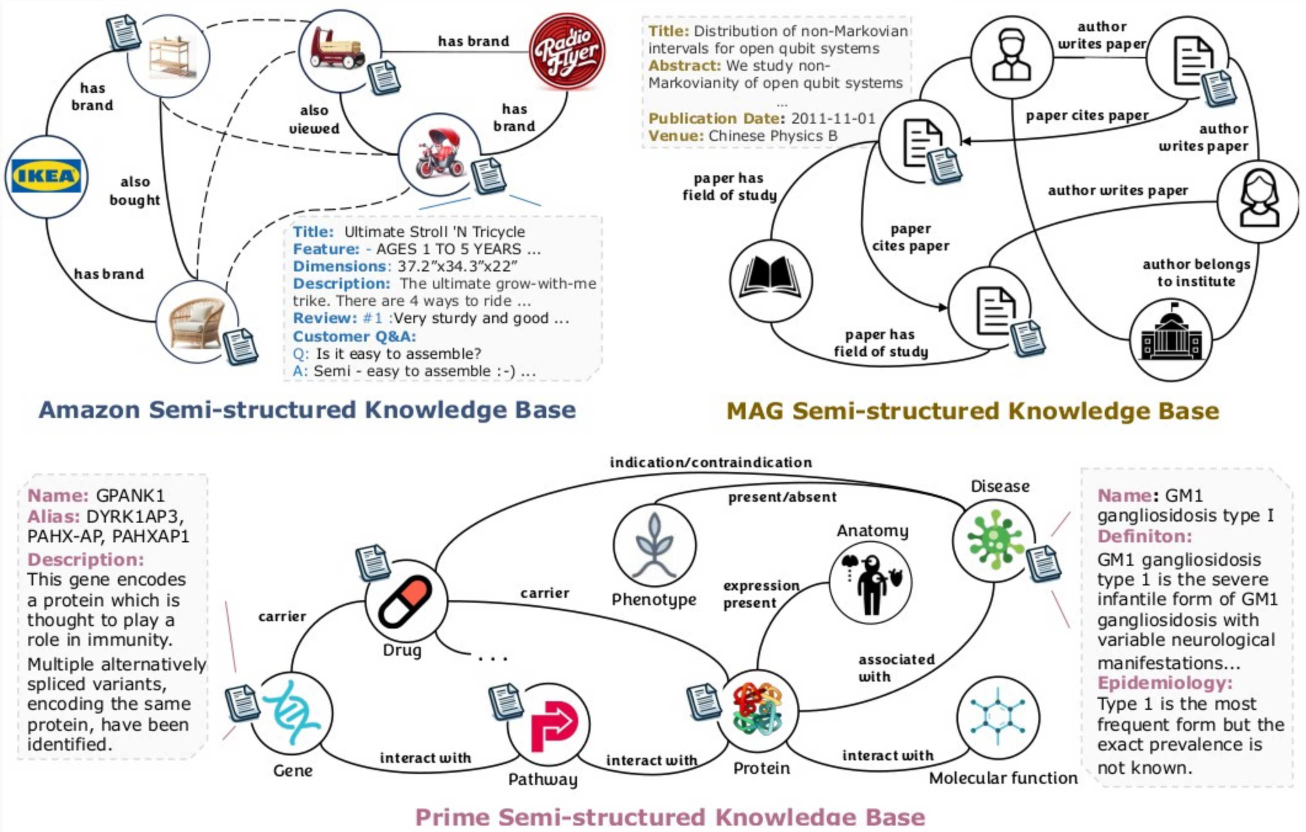
	Example query	Title of ground-truth items(s)
STaRK-Amazon	*Looking for durable Dart World brand dart flights that resist easy tearing. Any recommendations?*	<Amazon Standard Flights> <Dart World Broken Glass Flight> (+12 more)
	What are recommended scuba diving weights for experienced divers that would fit well with my Gorilla PRO XL waterproof bag?	<Sea Pearls Vinyl Coated Lace Thru Weight>
STaRK-MAG	*Search publications by Hao-Sheng Zeng on non-Markovian dynamics.*	<Distribution of non-Markovian intervals...> <Comparison between non-Markovian...>
	What are some nanofluid heat transfer research papers published by scholars from Philadelphia University?	<A Numerical Study on Convection Around A Suqare Cylinder using AL2O3-H2O Nanofluid>
STaRK-Prime	*Could you provide a list of investigational drugs that interact with genes or proteins active in the epididymal region?*	<(S)-3-phenyllactic Acid>, <Anisomycin>, <Puromycin>
	Search for diseases without known treatments and induce pruritus in pregnant women, potentially associated with Autoimmune.	<Intrahepatic Cholestasis>
	Please find pathways involving the POLR3D gene within nucleoplasm.	<RNA Polymerase III Chain Elongation>
	Which gene or protein associated with lichen amyloidosis can bind interleukin-31 to activate the PI3K/AKT and MAPK pathways?	<OSMR>, <IL31RA>

Table 1: Example queries from STaRK, which involve semi-structured information, where the relational and textual aspects are highlighted.

А что раньше?

- Поиск по тексту
- Генерация запросов к графу
- Генерация SQL
- Если отношения — то общие, из википедии
- Отношения википедии неявно есть в языке
- Корпоративные данные устроены иначе (наверное)

Датасеты



Title: Ultimate Stroll 'N Tricycle
Feature: - AGES 1 TO 5 YEARS ...
Dimensions: 37.2″x34.3″x22″
Description: The ultimate grow-with-me trike. There are 4 ways to ride ...
Review: #1 :Very sturdy and good ...
Customer Q&A:
Q: Is it easy to assemble?
A: Semi - easy to assemble :-) ...

Amazon Semi-structured Knowledge Base

Title: Distribution of non-Markovian intervals for open qubit systems
Abstract: We study non-Markovianity of open qubit systems
...
Publication Date: 2011-11-01
Venue: Chinese Physics B

MAG Semi-structured Knowledge Base

Name: GPANK1
Alias: DYRK1AP3, PAHX-AP, PAHXAP1
Description:
This gene encodes a protein which is thought to play a role in immunity.
Multiple alternatively spliced variants, encoding the same protein, have been identified.

Name: GM1 gangliosidosis type I
Definiton:
GM1 gangliosidosis type 1 is the severe infantile form of GM1 gangliosidosis with variable neurological manifestations...
Epidemiology:
Type 1 is the most frequent form but the exact prevalence is not known.

Prime Semi-structured Knowledge Base

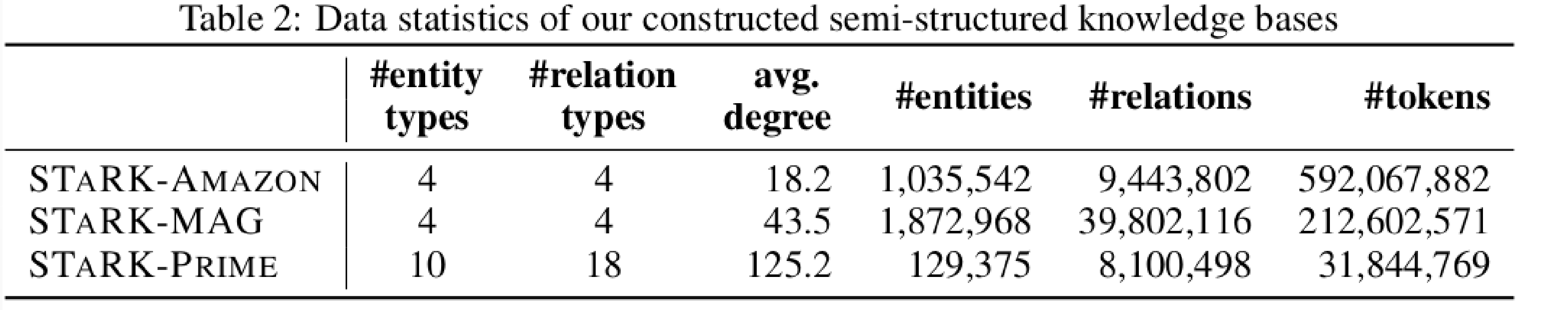


Table 2: Data statistics of our constructed semi-structured knowledge bases

	#entity types	#relation types	avg. degree	#entities	#relations	#tokens
STaRK-Amazon	4	4	18.2	1,035,542	9,443,802	592,067,882
STaRK-MAG	4	4	43.5	1,872,968	39,802,116	212,602,571
STaRK-Prime	10	18	125.2	129,375	8,100,498	31,844,769

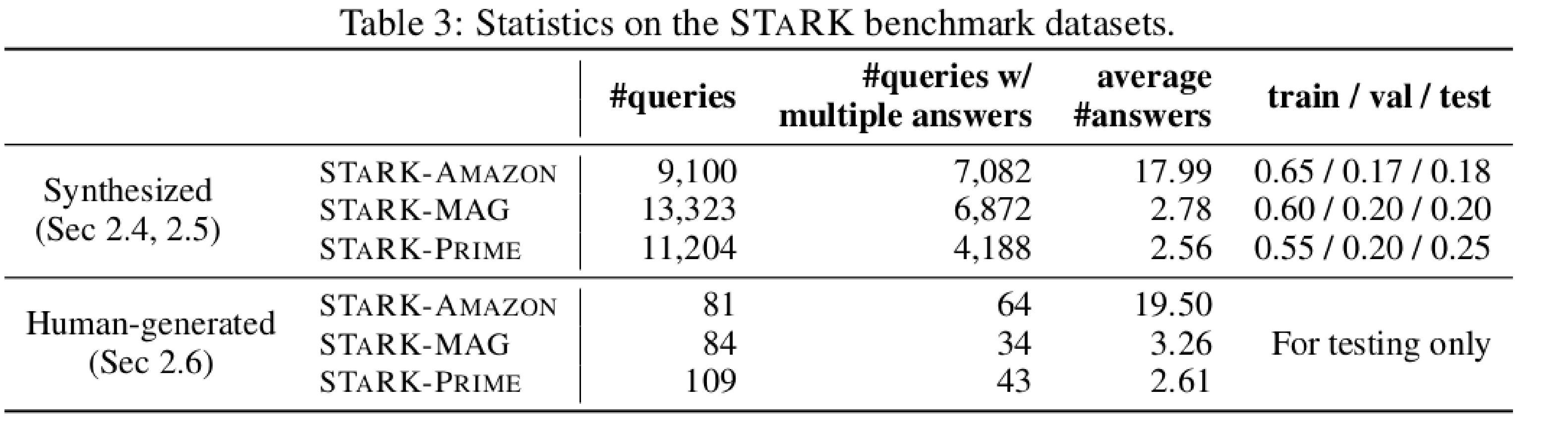


Table 3: Statistics on the STaRK benchmark datasets.

		#queries	#queries w/ multiple answers	average #answers	train / val / test
Synthesized (Sec 2.4, 2.5)	STaRK-Amazon	9,100	7,082	17.99	0.65 / 0.17 / 0.18
	STaRK-MAG	13,323	6,872	2.78	0.60 / 0.20 / 0.20
	STaRK-Prime	11,204	4,188	2.56	0.55 / 0.20 / 0.25
Human-generated (Sec 2.6)	STaRK-Amazon	81	64	19.50	For testing only
	STaRK-MAG	84	34	3.26	
	STaRK-Prime	109	43	2.61	

Генерация синтетики

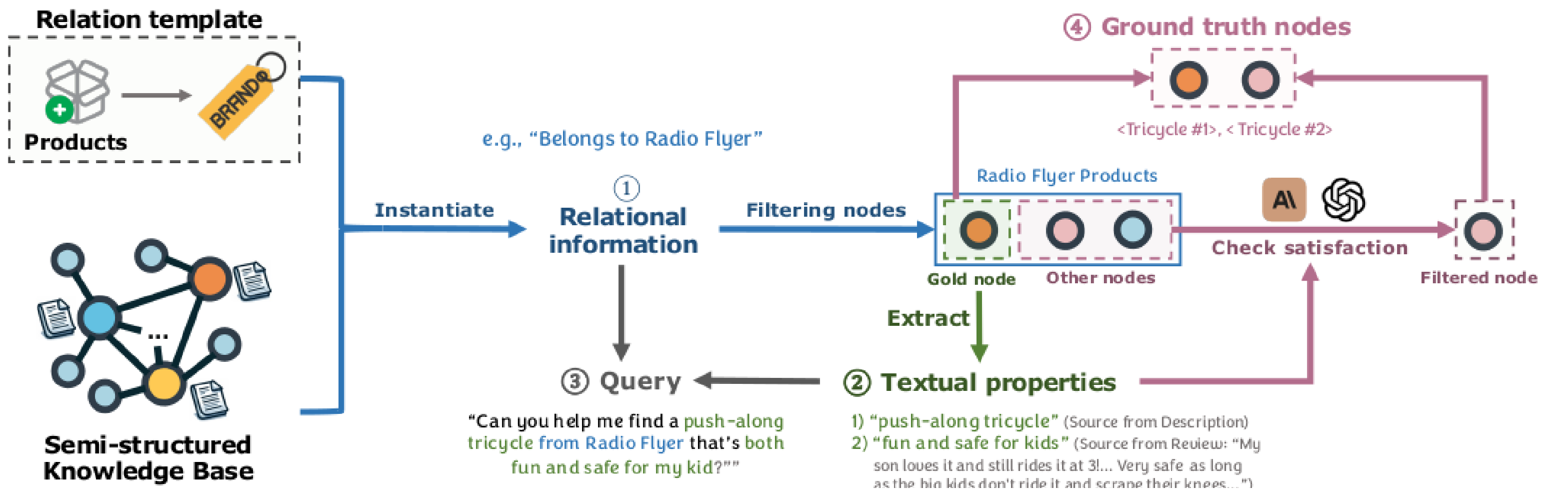
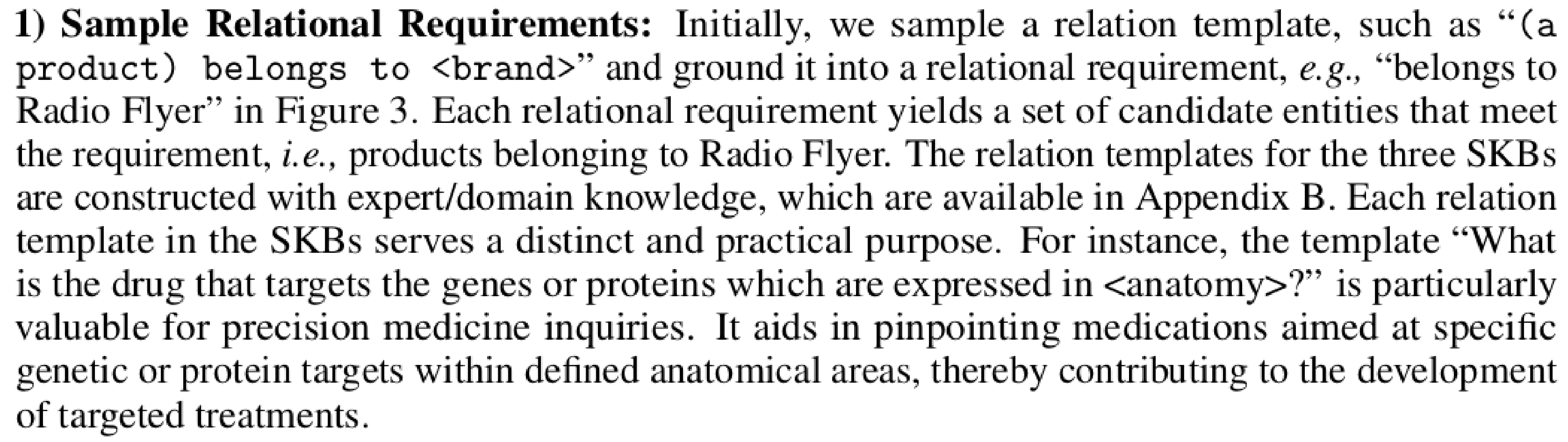
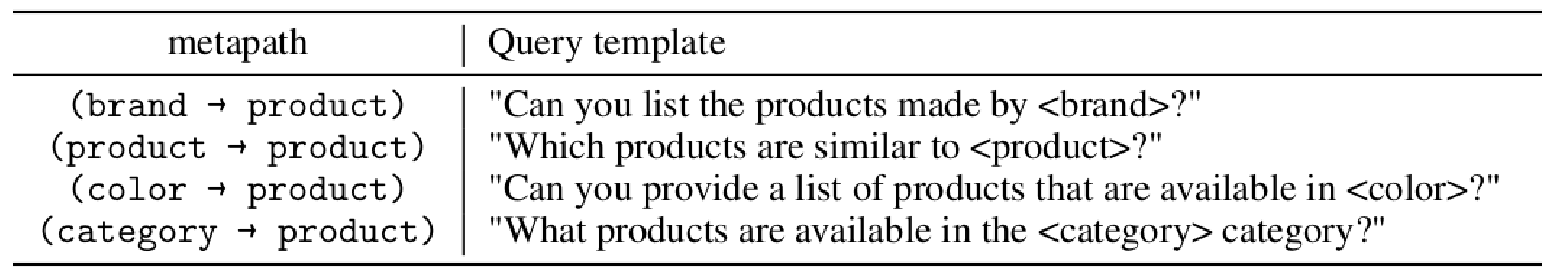

Relation template
Products
BRAND
Semi-structured Knowledge Base
Instantiate
e.g., "Belongs to Radio Flyer"
① Relational information
Filtering nodes
Radio Flyer Products
Gold node
Other nodes
④ Ground truth nodes
<Tricycle #1>, < Tricycle #2>
Check satisfaction
Filtered node
Extract
② Textual properties
1) "push-along tricycle" (Source from Description)
2) "fun and safe for kids" (Source from Review: "My son loves it and still rides it at 3!... Very safe as long as the big kids don't ride it and scrape their knees...")
③ Query
"Can you help me find a push-along tricycle from Radio Flyer that's both fun and safe for my kid?""

Figure 3: The process of constructing semi-structured retrieval datasets involves four main steps: 1) Sample Relational Requirement: Based on relational templates, sample a relational requirement on a SKB. 2) Extract Textual Properties: From a node that meets the relational requirement, extract relevant textual properties. 3) Combine Information: Merge the relational information and textual properties to form a natural-sounding query. 4) Construct Ground Truth Nodes: Check if nodes satisfy the textual properties using multiple language models to establish ground truth nodes.

Сэмплрируем из графа

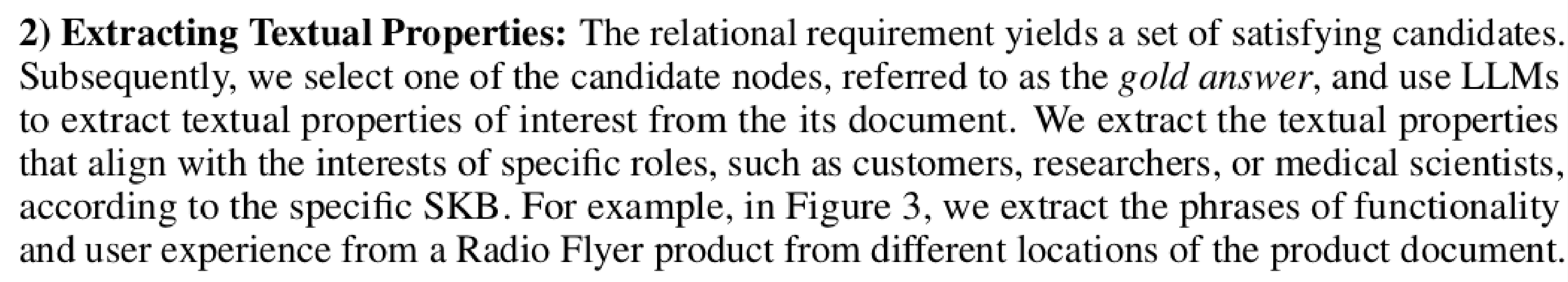


1) Sample Relational Requirements: Initially, we sample a relation template, such as "`(a product) belongs to <brand>`" and ground it into a relational requirement, *e.g.*, "belongs to Radio Flyer" in Figure 3. Each relational requirement yields a set of candidate entities that meet the requirement, *i.e.*, products belonging to Radio Flyer. The relation templates for the three SKBs are constructed with expert/domain knowledge, which are available in Appendix B. Each relation template in the SKBs serves a distinct and practical purpose. For instance, the template "What is the drug that targets the genes or proteins which are expressed in <anatomy>?" is particularly valuable for precision medicine inquiries. It aids in pinpointing medications aimed at specific genetic or protein targets within defined anatomical areas, thereby contributing to the development of targeted treatments.

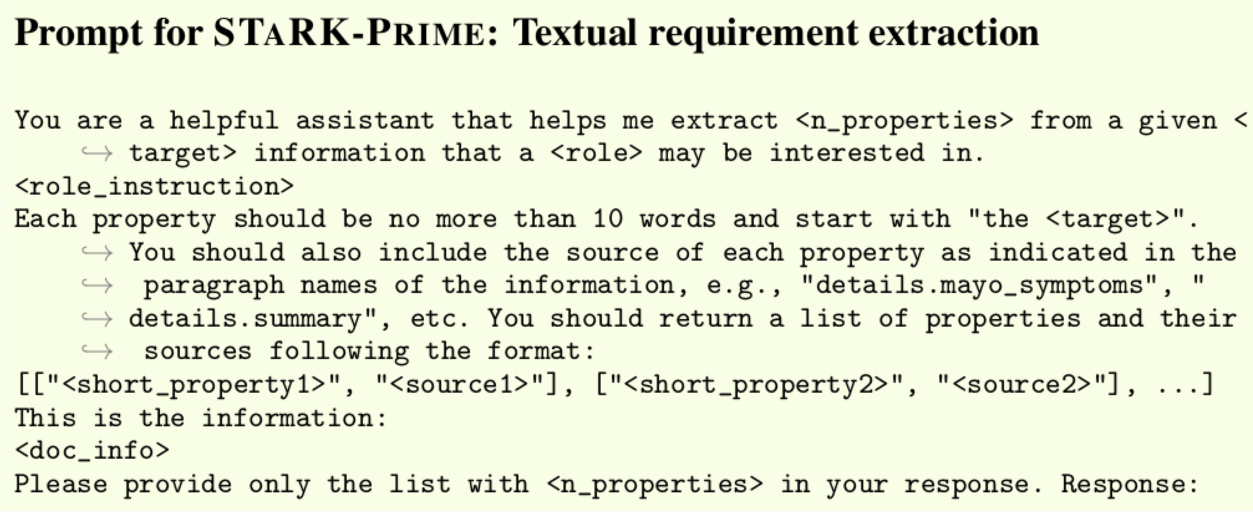


metapath	Query template
`(brand → product)`	"Can you list the products made by <brand>?"
`(product → product)`	"Which products are similar to <product>?"
`(color → product)`	"Can you provide a list of products that are available in <color>?"
`(category → product)`	"What products are available in the <category> category?"

Извлекаем свойства



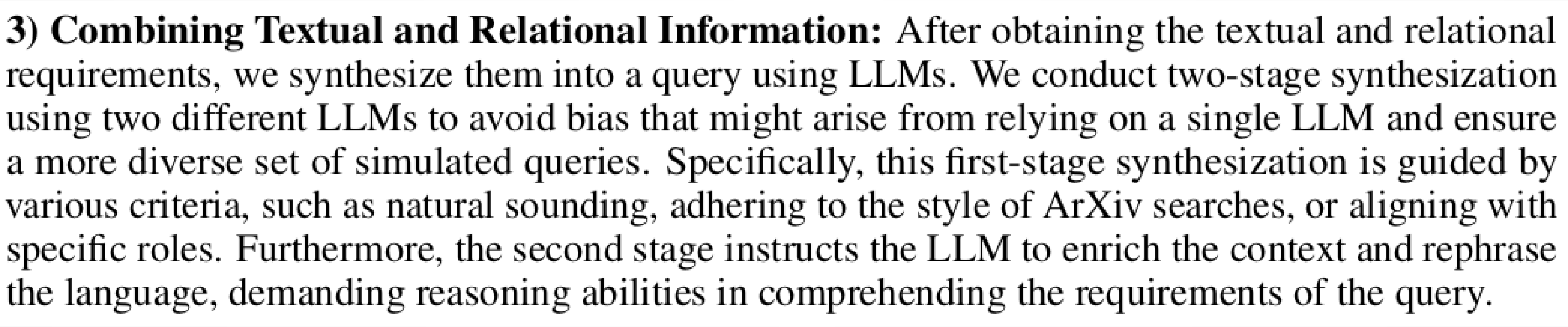
2) Extracting Textual Properties: The relational requirement yields a set of satisfying candidates. Subsequently, we select one of the candidate nodes, referred to as the *gold answer*, and use LLMs to extract textual properties of interest from the its document. We extract the textual properties that align with the interests of specific roles, such as customers, researchers, or medical scientists, according to the specific SKB. For example, in Figure 3, we extract the phrases of functionality and user experience from a Radio Flyer product from different locations of the product document.



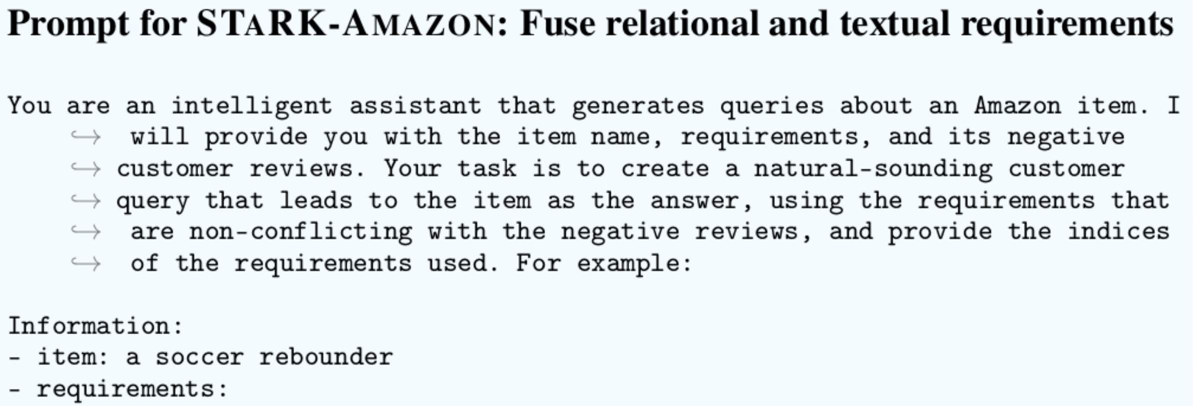
Prompt for STARK-PRIME: Textual requirement extraction

```
You are a helpful assistant that helps me extract <n_properties> from a given <
    ↪ target> information that a <role> may be interested in.
<role_instruction>
Each property should be no more than 10 words and start with "the <target>".
    ↪ You should also include the source of each property as indicated in the
    ↪  paragraph names of the information, e.g., "details.mayo_symptoms", "
    ↪ details.summary", etc. You should return a list of properties and their
    ↪  sources following the format:
[["<short_property1>", "<source1>"], ["<short_property2>", "<source2>"], ...]
This is the information:
<doc_info>
Please provide only the list with <n_properties> in your response. Response:
```

Собираем запрос



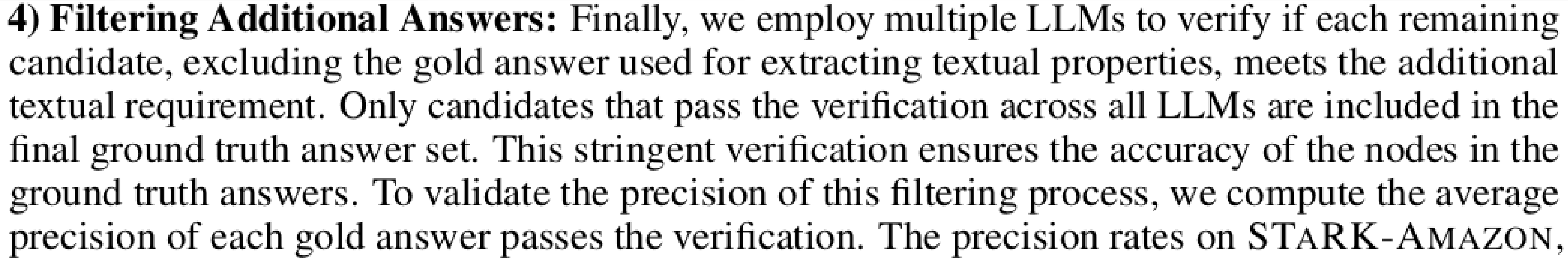
3) Combining Textual and Relational Information: After obtaining the textual and relational requirements, we synthesize them into a query using LLMs. We conduct two-stage synthesization using two different LLMs to avoid bias that might arise from relying on a single LLM and ensure a more diverse set of simulated queries. Specifically, this first-stage synthesization is guided by various criteria, such as natural sounding, adhering to the style of ArXiv searches, or aligning with specific roles. Furthermore, the second stage instructs the LLM to enrich the context and rephrase the language, demanding reasoning abilities in comprehending the requirements of the query.



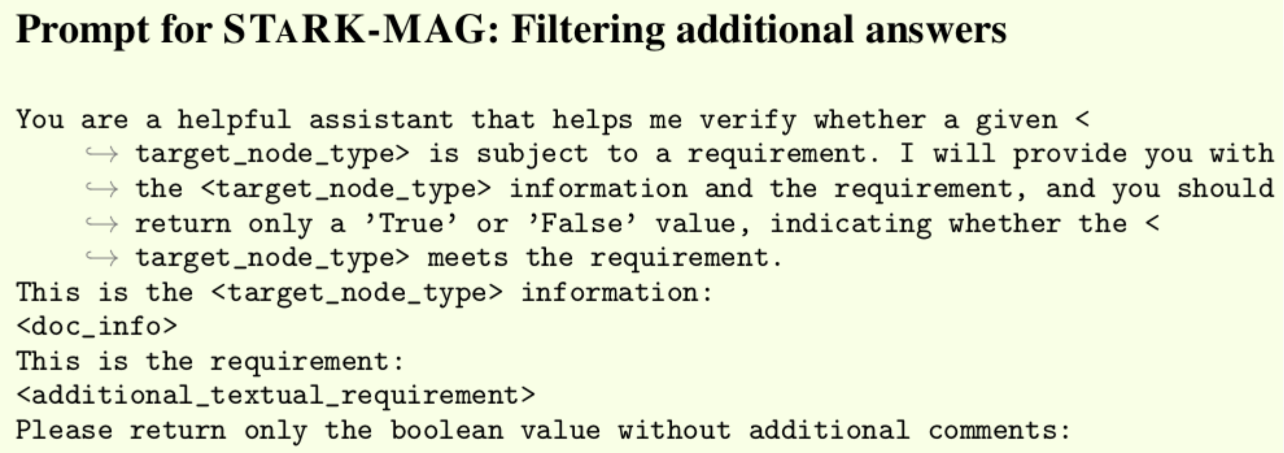
Prompt for STaRK-Amazon: Fuse relational and textual requirements

```
You are an intelligent assistant that generates queries about an Amazon item. I
    ↪  will provide you with the item name, requirements, and its negative
    ↪ customer reviews. Your task is to create a natural-sounding customer
    ↪ query that leads to the item as the answer, using the requirements that
    ↪  are non-conflicting with the negative reviews, and provide the indices
    ↪  of the requirements used. For example:

Information:
- item: a soccer rebounder
- requirements:
```

Выкидываем неоднозначности



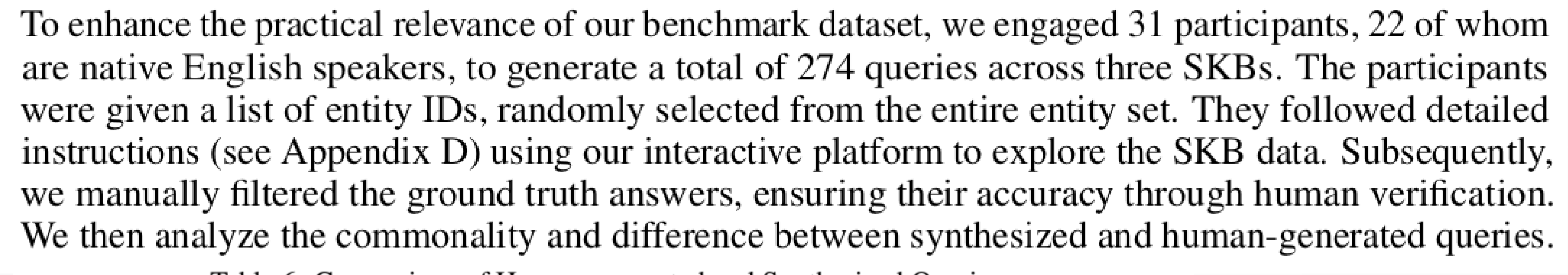
4) Filtering Additional Answers: Finally, we employ multiple LLMs to verify if each remaining candidate, excluding the gold answer used for extracting textual properties, meets the additional textual requirement. Only candidates that pass the verification across all LLMs are included in the final ground truth answer set. This stringent verification ensures the accuracy of the nodes in the ground truth answers. To validate the precision of this filtering process, we compute the average precision of each gold answer passes the verification. The precision rates on STaRK-Amazon,



Prompt for STaRK-MAG: Filtering additional answers

```
You are a helpful assistant that helps me verify whether a given <
    ↪ target_node_type> is subject to a requirement. I will provide you with
    ↪ the <target_node_type> information and the requirement, and you should
    ↪ return only a 'True' or 'False' value, indicating whether the <
    ↪ target_node_type> meets the requirement.
This is the <target_node_type> information:
<doc_info>
This is the requirement:
<additional_textual_requirement>
Please return only the boolean value without additional comments:
```

Люди тоже участвовали



To enhance the practical relevance of our benchmark dataset, we engaged 31 participants, 22 of whom are native English speakers, to generate a total of 274 queries across three SKBs. The participants were given a list of entity IDs, randomly selected from the entire entity set. They followed detailed instructions (see Appendix D) using our interactive platform to explore the SKB data. Subsequently, we manually filtered the ground truth answers, ensuring their accuracy through human verification. We then analyze the commonality and difference between synthesized and human-generated queries.

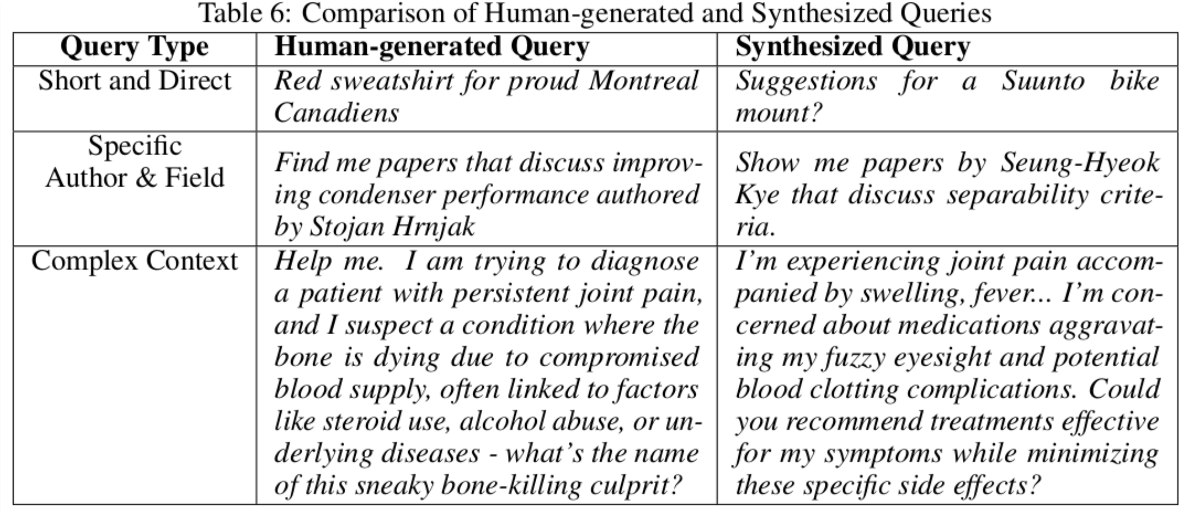
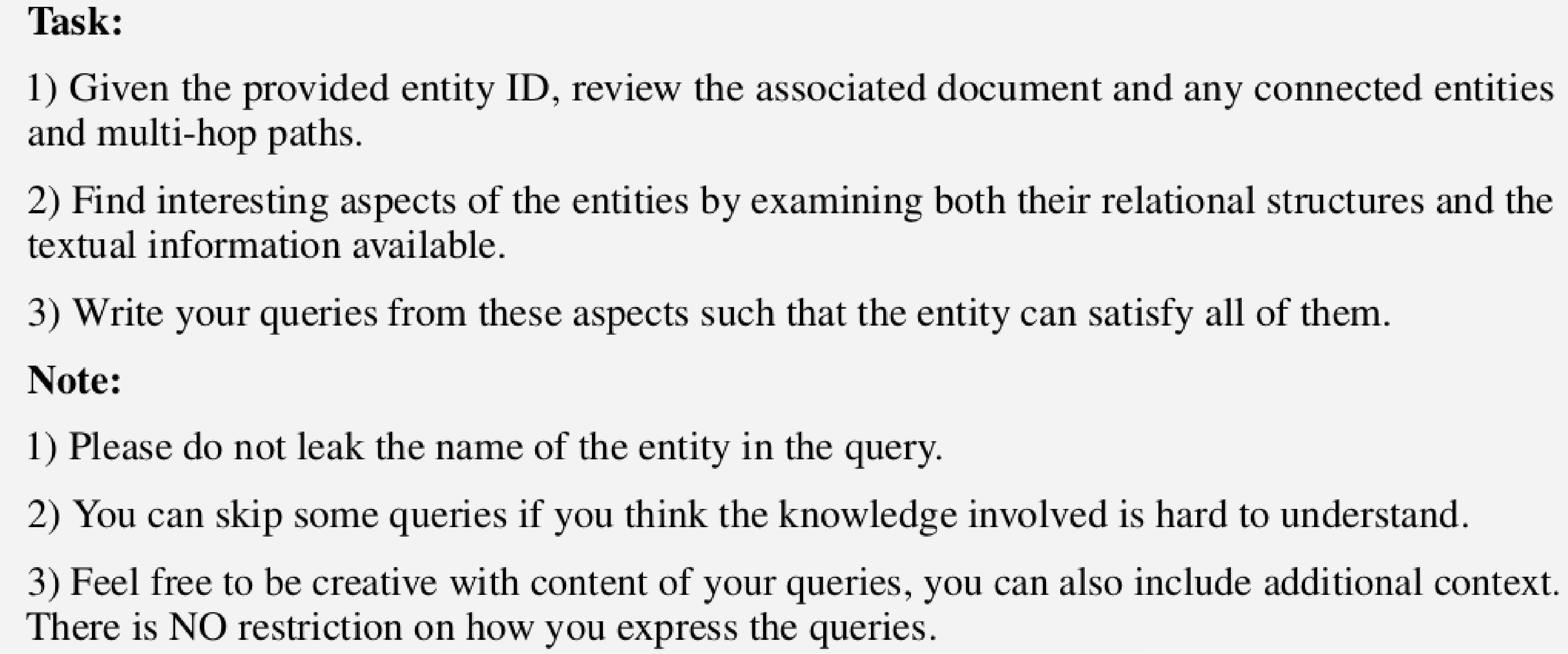


Table 6: Comparison of Human-generated and Synthesized Queries

Query Type	**Human-generated Query**	**Synthesized Query**
Short and Direct	*Red sweatshirt for proud Montreal Canadiens*	*Suggestions for a Suunto bike mount?*
Specific Author & Field	*Find me papers that discuss improving condenser performance authored by Stojan Hrnjak*	*Show me papers by Seung-Hyeok Kye that discuss separability criteria.*
Complex Context	*Help me. I am trying to diagnose a patient with persistent joint pain, and I suspect a condition where the bone is dying due to compromised blood supply, often linked to factors like steroid use, alcohol abuse, or underlying diseases - what's the name of this sneaky bone-killing culprit?*	*I'm experiencing joint pain accompanied by swelling, fever... I'm concerned about medications aggravating my fuzzy eyesight and potential blood clotting complications. Could you recommend treatments effective for my symptoms while minimizing these specific side effects?*

Промпт для людей



Task:

1) Given the provided entity ID, review the associated document and any connected entities and multi-hop paths.

2) Find interesting aspects of the entities by examining both their relational structures and the textual information available.

3) Write your queries from these aspects such that the entity can satisfy all of them.

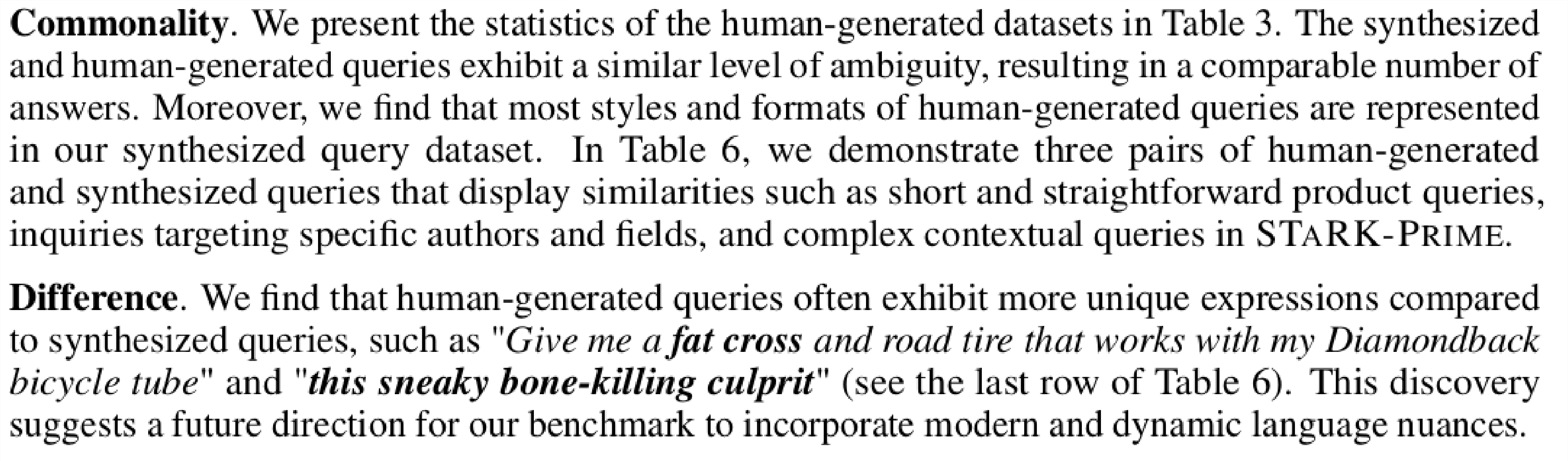
Note:

1) Please do not leak the name of the entity in the query.

2) You can skip some queries if you think the knowledge involved is hard to understand.

3) Feel free to be creative with content of your queries, you can also include additional context. There is NO restriction on how you express the queries.

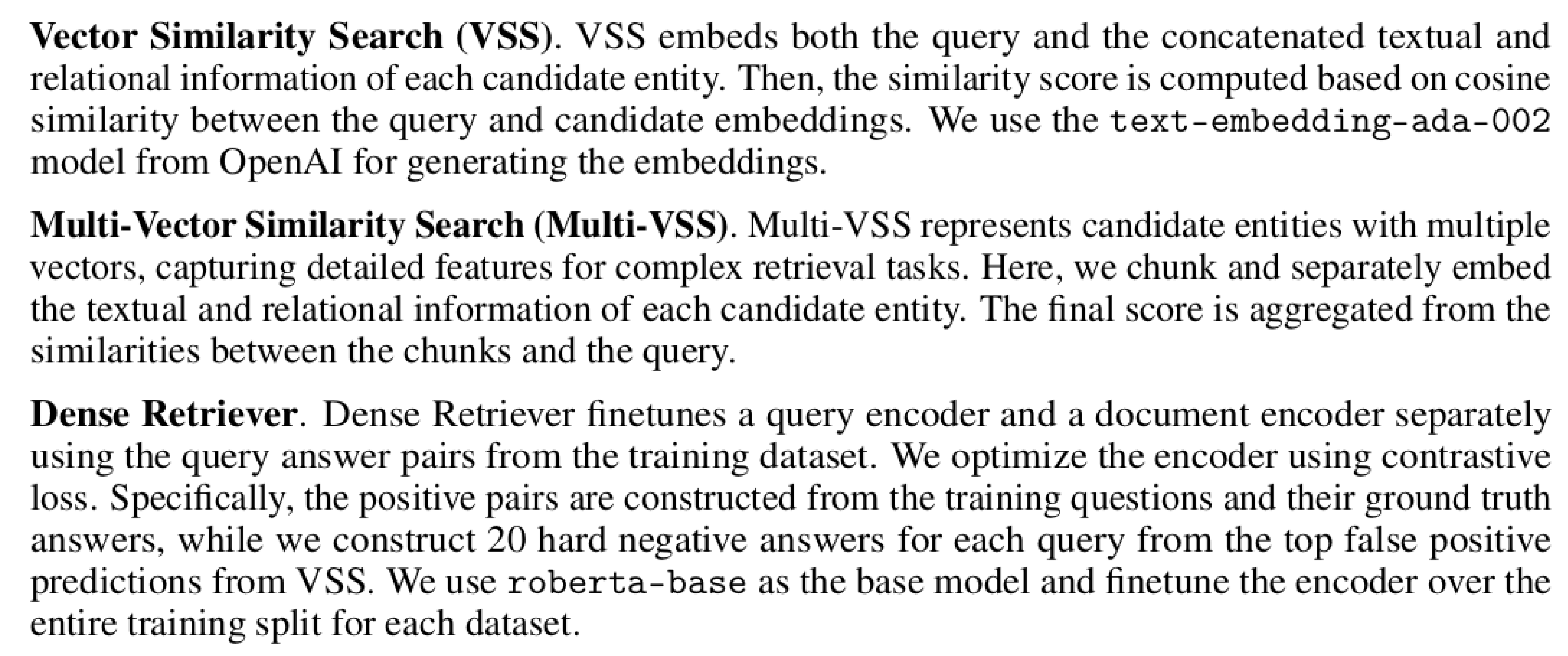
Люди против машин



Commonality. We present the statistics of the human-generated datasets in Table 3. The synthesized and human-generated queries exhibit a similar level of ambiguity, resulting in a comparable number of answers. Moreover, we find that most styles and formats of human-generated queries are represented in our synthesized query dataset. In Table 6, we demonstrate three pairs of human-generated and synthesized queries that display similarities such as short and straightforward product queries, inquiries targeting specific authors and fields, and complex contextual queries in STaRK-Prime.

Difference. We find that human-generated queries often exhibit more unique expressions compared to synthesized queries, such as "*Give me a* ***fat cross*** *and road tire that works with my Diamondback bicycle tube*" and "***this sneaky bone-killing culprit***" (see the last row of Table 6). This discovery suggests a future direction for our benchmark to incorporate modern and dynamic language nuances.

Подопытные модели

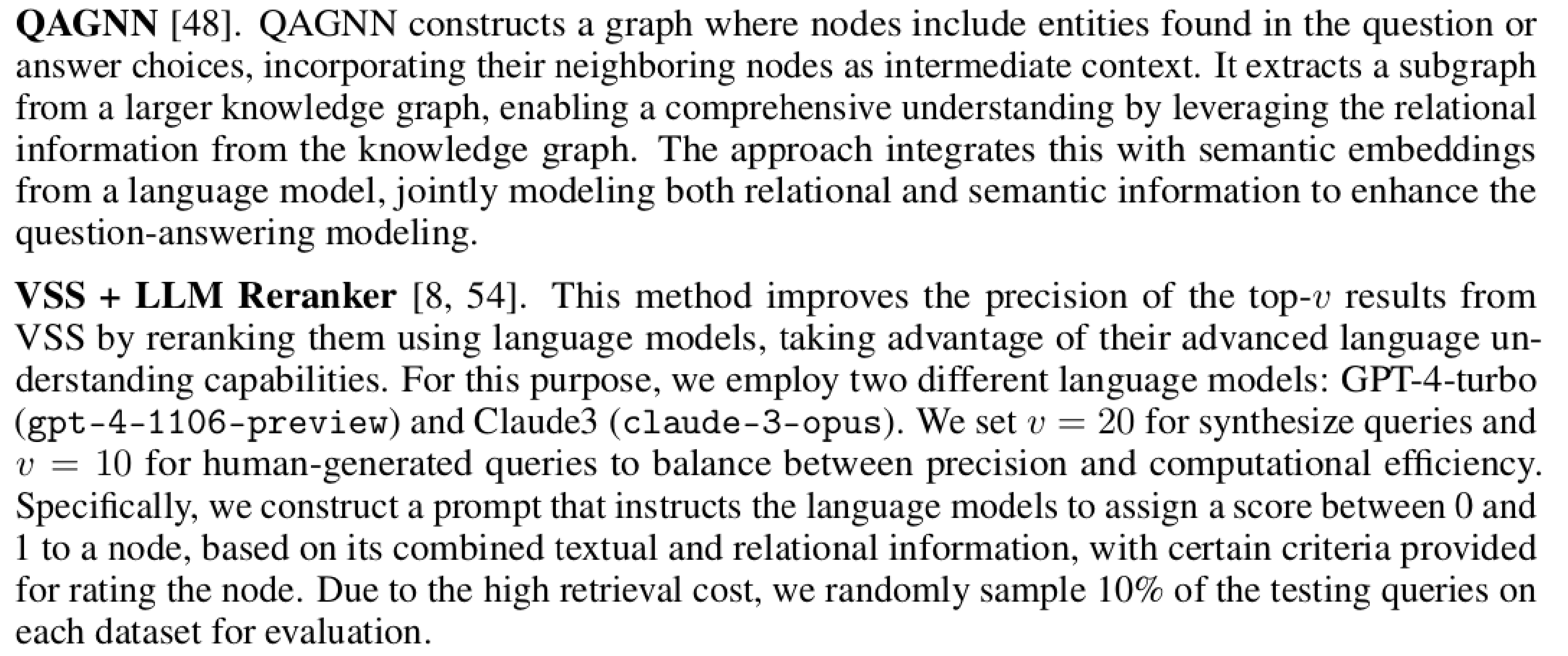


Vector Similarity Search (VSS). VSS embeds both the query and the concatenated textual and relational information of each candidate entity. Then, the similarity score is computed based on cosine similarity between the query and candidate embeddings. We use the `text-embedding-ada-002` model from OpenAI for generating the embeddings.

Multi-Vector Similarity Search (Multi-VSS). Multi-VSS represents candidate entities with multiple vectors, capturing detailed features for complex retrieval tasks. Here, we chunk and separately embed the textual and relational information of each candidate entity. The final score is aggregated from the similarities between the chunks and the query.

Dense Retriever. Dense Retriever finetunes a query encoder and a document encoder separately using the query answer pairs from the training dataset. We optimize the encoder using contrastive loss. Specifically, the positive pairs are constructed from the training questions and their ground truth answers, while we construct 20 hard negative answers for each query from the top false positive predictions from VSS. We use `roberta-base` as the base model and finetune the encoder over the entire training split for each dataset.

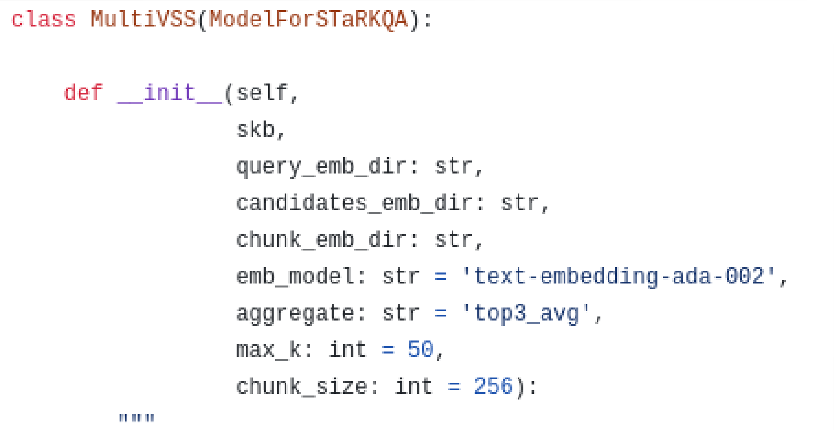
Еще подопытные



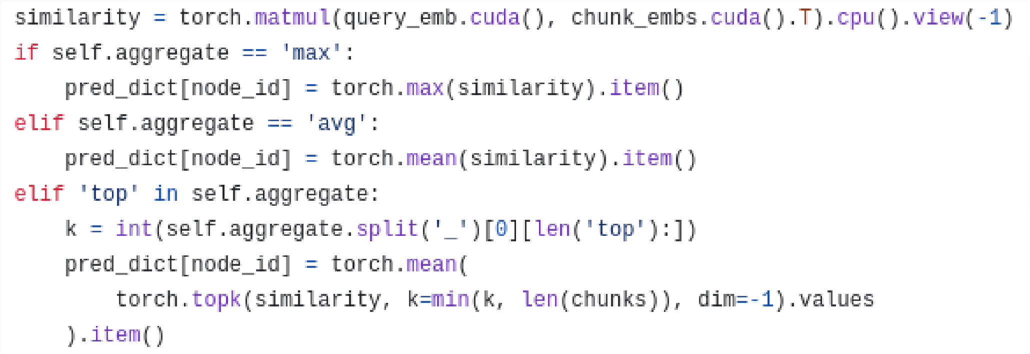
QAGNN [48]. QAGNN constructs a graph where nodes include entities found in the question or answer choices, incorporating their neighboring nodes as intermediate context. It extracts a subgraph from a larger knowledge graph, enabling a comprehensive understanding by leveraging the relational information from the knowledge graph. The approach integrates this with semantic embeddings from a language model, jointly modeling both relational and semantic information to enhance the question-answering modeling.

VSS + LLM Reranker [8, 54]. This method improves the precision of the top-v results from VSS by reranking them using language models, taking advantage of their advanced language understanding capabilities. For this purpose, we employ two different language models: GPT-4-turbo (`gpt-4-1106-preview`) and Claude3 (`claude-3-opus`). We set $v = 20$ for synthesize queries and $v = 10$ for human-generated queries to balance between precision and computational efficiency. Specifically, we construct a prompt that instructs the language models to assign a score between 0 and 1 to a node, based on its combined textual and relational information, with certain criteria provided for rating the node. Due to the high retrieval cost, we randomly sample 10% of the testing queries on each dataset for evaluation.

Mutli-VSS



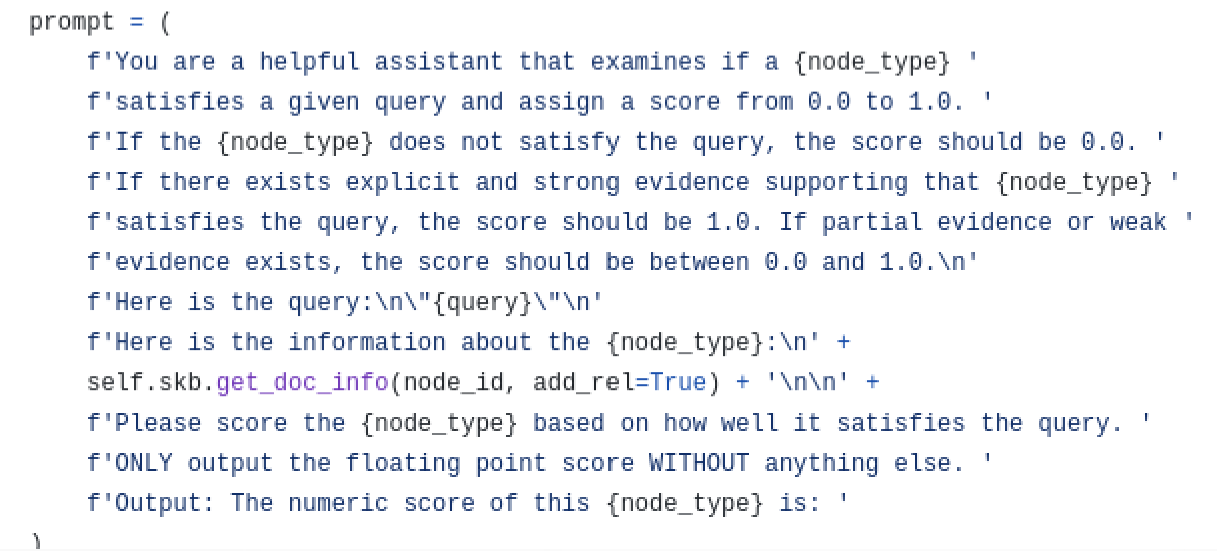
```python
class MultiVSS(ModelForSTaRKQA):

    def __init__(self,
                 skb,
                 query_emb_dir: str,
                 candidates_emb_dir: str,
                 chunk_emb_dir: str,
                 emb_model: str = 'text-embedding-ada-002',
                 aggregate: str = 'top3_avg',
                 max_k: int = 50,
                 chunk_size: int = 256):
        """
```



```python
similarity = torch.matmul(query_emb.cuda(), chunk_embs.cuda().T).cpu().view(-1)
if self.aggregate == 'max':
    pred_dict[node_id] = torch.max(similarity).item()
elif self.aggregate == 'avg':
    pred_dict[node_id] = torch.mean(similarity).item()
elif 'top' in self.aggregate:
    k = int(self.aggregate.split('_')[0][len('top'):])
    pred_dict[node_id] = torch.mean(
        torch.topk(similarity, k=min(k, len(chunks)), dim=-1).values
    ).item()
```

https://github.com/snap-stanford/stark

VSS+LLM Reranker



```
prompt = (
    f'You are a helpful assistant that examines if a {node_type} '
    f'satisfies a given query and assign a score from 0.0 to 1.0. '
    f'If the {node_type} does not satisfy the query, the score should be 0.0. '
    f'If there exists explicit and strong evidence supporting that {node_type} '
    f'satisfies the query, the score should be 1.0. If partial evidence or weak '
    f'evidence exists, the score should be between 0.0 and 1.0.\n'
    f'Here is the query:\n\"{query}\"\n'
    f'Here is the information about the {node_type}:\n' +
    self.skb.get_doc_info(node_id, add_rel=True) + '\n\n' +
    f'Please score the {node_type} based on how well it satisfies the query. '
    f'ONLY output the floating point score WITHOUT anything else. '
    f'Output: The numeric score of this {node_type} is: '
)
```

Товары находит, статьи нет

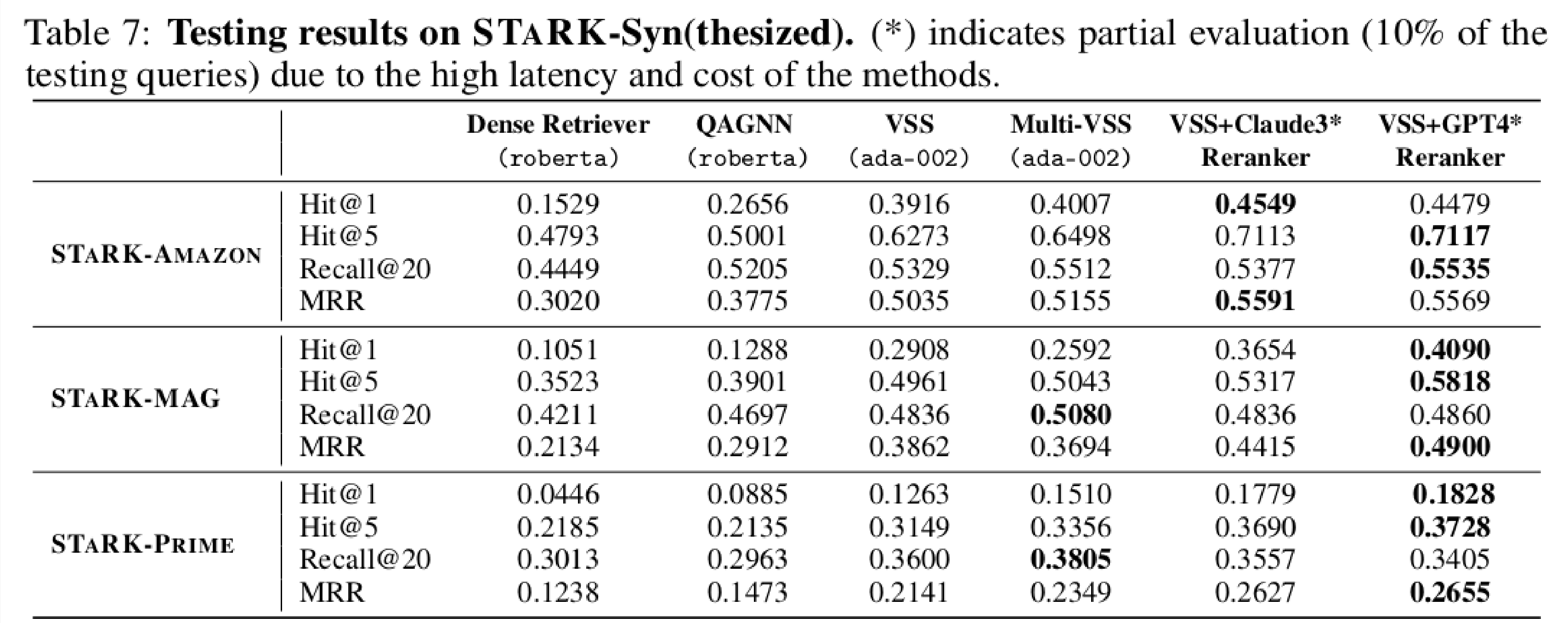


Table 7: **Testing results on STaRK-Syn(thesized).** (*) indicates partial evaluation (10% of the testing queries) due to the high latency and cost of the methods.

		Dense Retriever	**QAGNN**	**VSS**	**Multi-VSS**	**VSS+Claude3***	**VSS+GPT4***
		(roberta)	(roberta)	(ada-002)	(ada-002)	**Reranker**	**Reranker**
STaRK-Amazon	Hit@1	0.1529	0.2656	0.3916	0.4007	**0.4549**	0.4479
	Hit@5	0.4793	0.5001	0.6273	0.6498	0.7113	**0.7117**
	Recall@20	0.4449	0.5205	0.5329	0.5512	0.5377	**0.5535**
	MRR	0.3020	0.3775	0.5035	0.5155	**0.5591**	0.5569
STaRK-MAG	Hit@1	0.1051	0.1288	0.2908	0.2592	0.3654	**0.4090**
	Hit@5	0.3523	0.3901	0.4961	0.5043	0.5317	**0.5818**
	Recall@20	0.4211	0.4697	0.4836	**0.5080**	0.4836	0.4860
	MRR	0.2134	0.2912	0.3862	0.3694	0.4415	**0.4900**
STaRK-Prime	Hit@1	0.0446	0.0885	0.1263	0.1510	0.1779	**0.1828**
	Hit@5	0.2185	0.2135	0.3149	0.3356	0.3690	**0.3728**
	Recall@20	0.3013	0.2963	0.3600	**0.3805**	0.3557	0.3405
	MRR	0.1238	0.1473	0.2141	0.2349	0.2627	**0.2655**

MRR: Mean Reciprocal Rank

Человеческие запросы

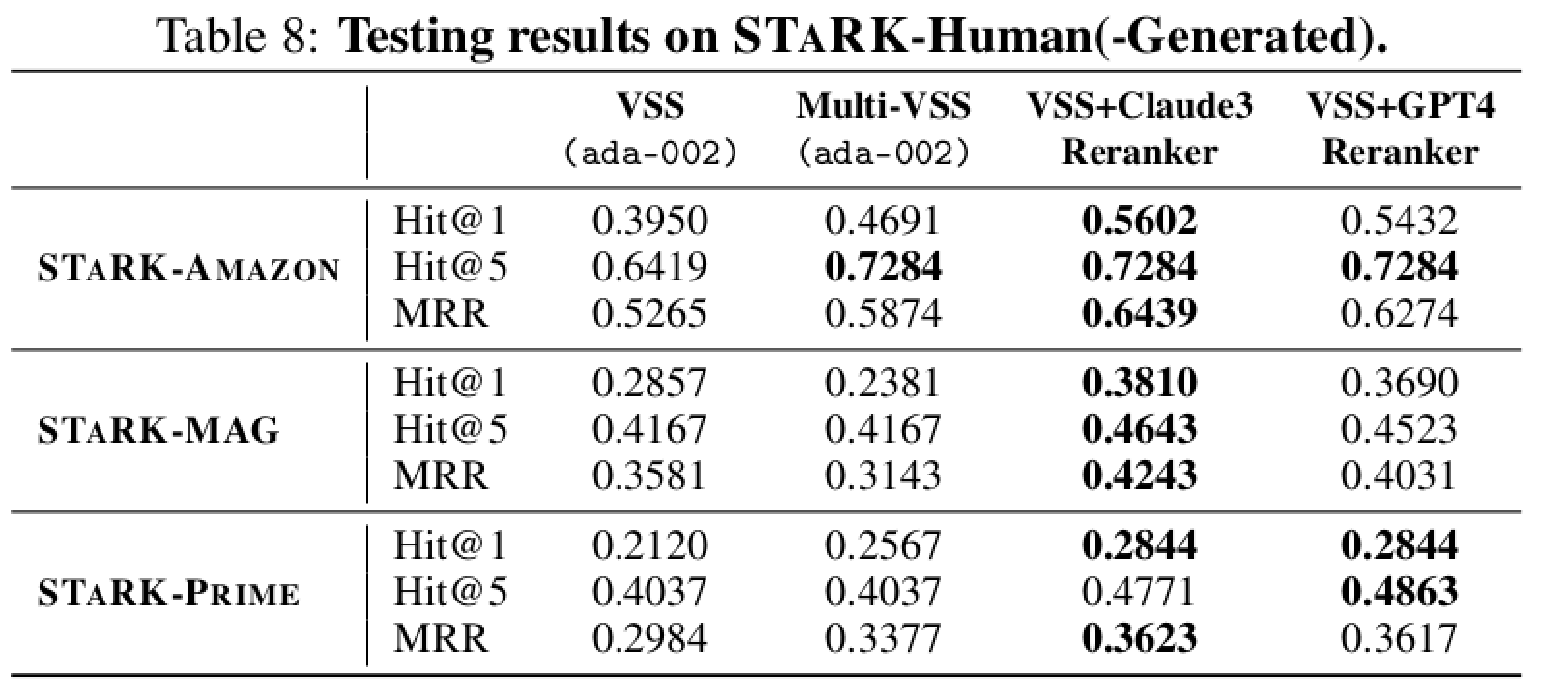


Table 8: **Testing results on STaRK-Human(-Generated).**

		VSS	**Multi-VSS**	**VSS+Claude3**	**VSS+GPT4**
		(ada-002)	(ada-002)	**Reranker**	**Reranker**
STaRK-Amazon	Hit@1	0.3950	0.4691	**0.5602**	0.5432
	Hit@5	0.6419	**0.7284**	**0.7284**	**0.7284**
	MRR	0.5265	0.5874	**0.6439**	0.6274
STaRK-MAG	Hit@1	0.2857	0.2381	**0.3810**	0.3690
	Hit@5	0.4167	0.4167	**0.4643**	0.4523
	MRR	0.3581	0.3143	**0.4243**	0.4031
STaRK-Prime	Hit@1	0.2120	0.2567	**0.2844**	**0.2844**
	Hit@5	0.4037	0.4037	0.4771	**0.4863**
	MRR	0.2984	0.3377	**0.3623**	0.3617

Притормозим

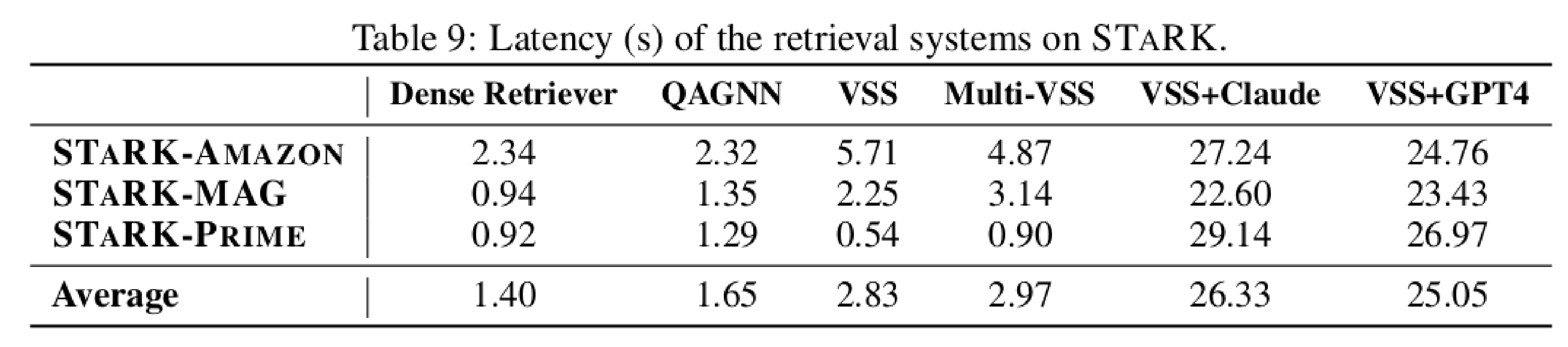
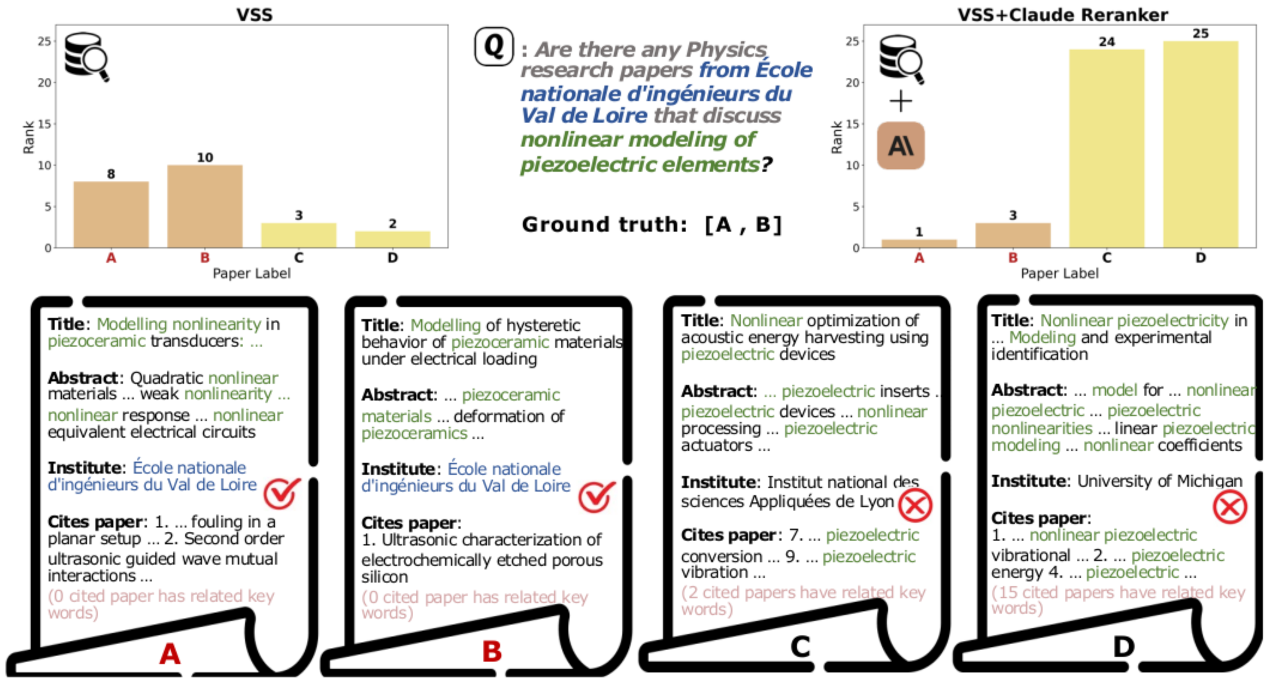


Table 9: Latency (s) of the retrieval systems on STaRK.

	Dense Retriever	QAGNN	VSS	Multi-VSS	VSS+Claude	VSS+GPT4
STaRK-Amazon	2.34	2.32	5.71	4.87	27.24	24.76
STaRK-MAG	0.94	1.35	2.25	3.14	22.60	23.43
STaRK-Prime	0.92	1.29	0.54	0.90	29.14	26.97
Average	1.40	1.65	2.83	2.97	26.33	25.05

Переранжирование помогает



Q : *Are there any Physics research papers from École nationale d'ingénieurs du Val de Loire that discuss nonlinear modeling of piezoelectric elements?*

Ground truth: [A , B]

A

Title: Modelling nonlinearity in piezoceramic transducers: ...

Abstract: Quadratic nonlinear materials ... weak nonlinearity ... nonlinear response ... nonlinear equivalent electrical circuits

Institute: École nationale d'ingénieurs du Val de Loire ✓

Cites paper: 1. ... fouling in a planar setup ... 2. Second order ultrasonic guided wave mutual interactions ...

(0 cited paper has related key words)

B

Title: Modelling of hysteretic behavior of piezoceramic materials under electrical loading

Abstract: ... piezoceramic materials ... deformation of piezoceramics ...

Institute: École nationale d'ingénieurs du Val de Loire ✓

Cites paper:
1. Ultrasonic characterization of electrochemically etched porous silicon

(0 cited paper has related key words)

C

Title: Nonlinear optimization of acoustic energy harvesting using piezoelectric devices

Abstract: ... piezoelectric inserts ... piezoelectric devices ... nonlinear processing ... piezoelectric actuators ...

Institute: Institut national des sciences Appliquées de Lyon ⊗

Cites paper: 7. ... piezoelectric conversion ... 9. ... piezoelectric vibration ...

(2 cited papers have related key words)

D

Title: Nonlinear piezoelectricity in ... Modeling and experimental identification

Abstract: ... model for ... nonlinear piezoelectric ... piezoelectric nonlinearities ... linear piezoelectric modeling ... nonlinear coefficients

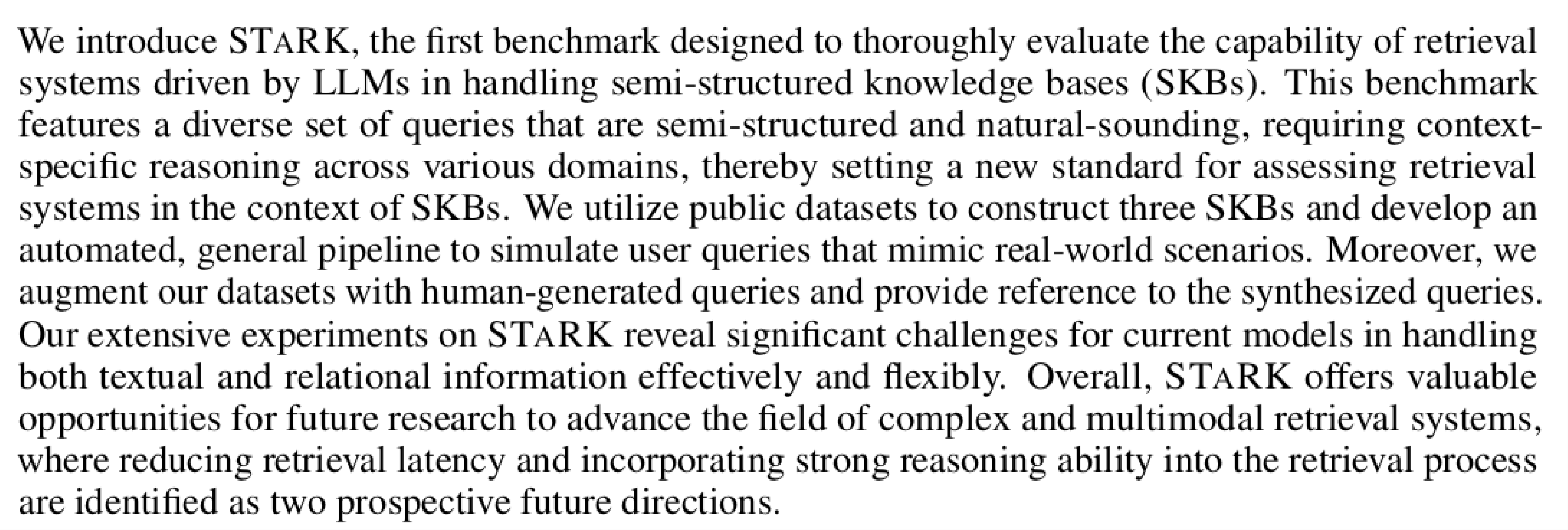
Institute: University of Michigan ⊗

Cites paper:
1. ... nonlinear piezoelectric vibrational ... 2. ... piezoelectric energy 4. ... piezoelectric ...

(15 cited papers have related key words)

Figure 6: A case study on STaRK-MAG, where VSS mistakenly ranks non-ground truth papers C and D higher due to the repeated key words in the relational information "cites paper." After reranking with Claude, it correctly prioritizes the ground truth papers A and B. This correction is attributed to the more accurate reasoning and analysis of the combined textual and relational information.

Итого



We introduce STaRK, the first benchmark designed to thoroughly evaluate the capability of retrieval systems driven by LLMs in handling semi-structured knowledge bases (SKBs). This benchmark features a diverse set of queries that are semi-structured and natural-sounding, requiring context-specific reasoning across various domains, thereby setting a new standard for assessing retrieval systems in the context of SKBs. We utilize public datasets to construct three SKBs and develop an automated, general pipeline to simulate user queries that mimic real-world scenarios. Moreover, we augment our datasets with human-generated queries and provide reference to the synthesized queries. Our extensive experiments on STaRK reveal significant challenges for current models in handling both textual and relational information effectively and flexibly. Overall, STaRK offers valuable opportunities for future research to advance the field of complex and multimodal retrieval systems, where reducing retrieval latency and incorporating strong reasoning ability into the retrieval process are identified as two prospective future directions.

А на самом деле

- Предложен хороший (простой, надежный, верифицируемый) фреймворк для генерации тестовых и обучающих датасетов на ваших данных, если они содержат текстовые данные и графы
- Предложена методика оценки качества синтетических данных
- Элегантно отбрасывают неоднозначности