Can ChatGPT Replace Traditional KBQA Models? An In-depth Analysis of the Question Answering Performance of the GPT LLM Family

Yiming Tan*, Dehai Min*, Yu Li, Wenbo Li,
Nan Hu, Yongrui Chen, Guilin Qi

Southeast University

11.7.2023

Outline

1. Background and research objectives
2. Previous works and findings
3. The Q&A evaluation framework
4. Experiments and findings
5. Conclusion

Part 1

Background and Research Objectives

Large language models like GPT family contain vast amounts of knowledge and support answering questions posed by users using their own included knowledge.

Question: Can large language models replace the traditional KBQA model ?

Research objective: To evaluate the effectiveness of large language models, represented by the GPT family, when used as self-referential knowledge graphs in answering complex open-domain questions.

Part 2

Previous works and findings

Previous findings:

- ChatGPT tends to be a lazy reasoner and performs poorly in inductive reasoning tasks. (Bang et. al, A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity, 2023)

- ChatGPT exhibits lower consistency in its question-answering results compared to traditional KBQA models. (Omar et. al, Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots, 2023)

Previous work

Benchmark

Natural Questions

XSUM

IMDB

MS MARCO

CivilComments

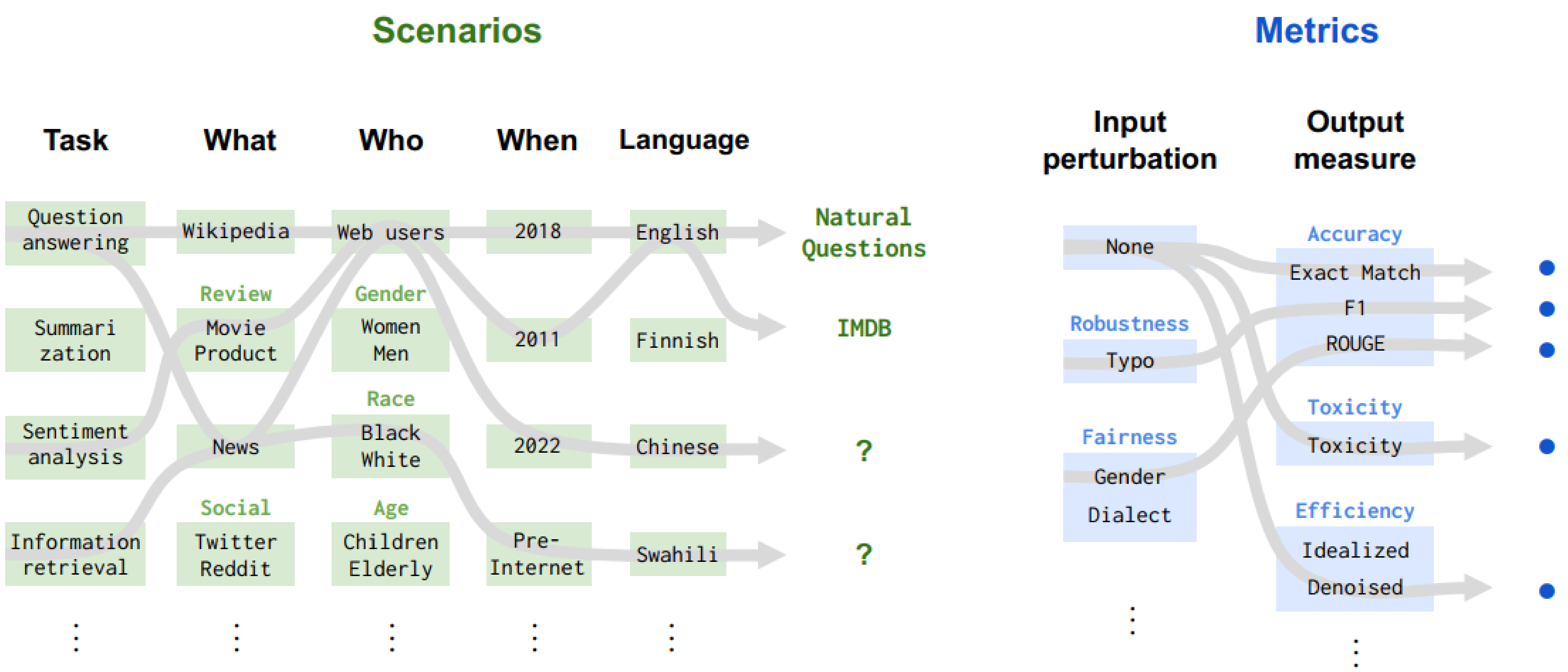
WikiText-103

WebNLG

ANLI

⋮

HELM



(Liang et.al, Holistic Evaluation of Language Models, 2022)

CheckList[4] Black-box testing

1. Minimum Functionality Test
   - Testing the model's various fundamental

2. INVariance Test
   - Making multiple input modifications while keeping the main features unchanged, observe if the model can maintain output consistency.

3. DIRectional Expectation Test
   - Introducing expected input modifications to observe whether the model produces the anticipated results.

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

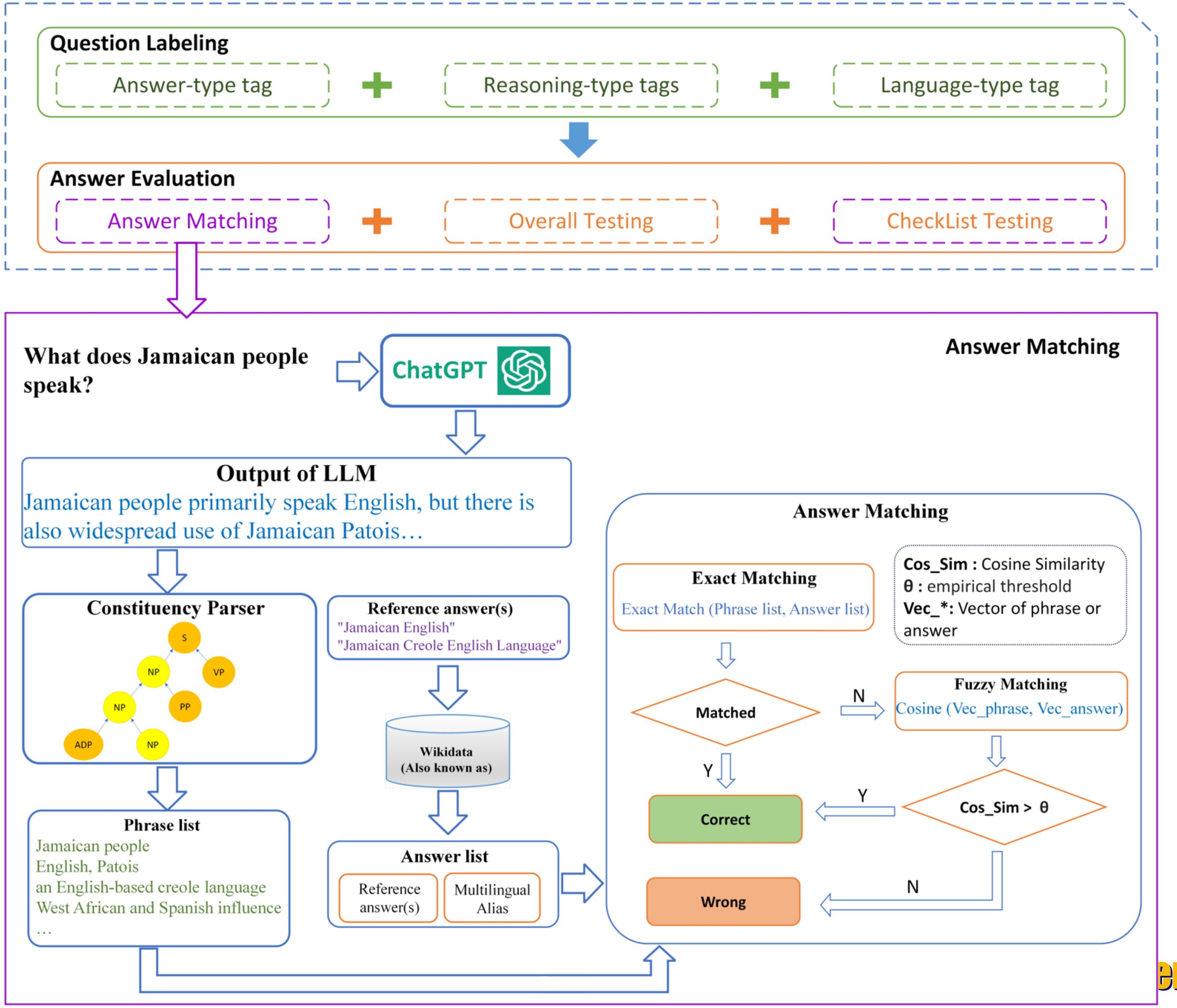
Test case	Expected	Predicted	Pass?
A Testing Negation with *MFT* Labels: negative, positive, neutral			
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing NER with *INV* Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos / neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral / neg	X
...			
Failure rate = 20.8%			
C Testing Vocabulary with *DIR* Sentiment monotonic decreasing (↓)			
@AmericanAir service wasn't great. You are lame.	↓	neg / neutral	X
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg / neutral	X
...			
Failure rate = 34.6%			

(Ribeiro et. al, Beyond Accuracy: Behavioral Testing of NLP models with CheckList, 2020)

Part 3

The Q&A evaluation framework

The Q&A evaluation framework



Question Labeling:

The three labels "Answer-Type," "Reasoning-Type," and "Language-Type" are set to uniformly describe the characteristics of questions originating from different KBQA data sets.

Answer Evaluation:

Answer Matching:
Exact matching (EM) + Fuzzy matching

Overall Testing: Assessment of QA Performance for GPT LLM.

CheckList Testing: Testing the Consistency and Robustness of GPT LLM as a Question-Answering System

Question Labeling – Feature tags

1. Source of feature labels:

Answer Type: From answer types in existing KBQA datasets.

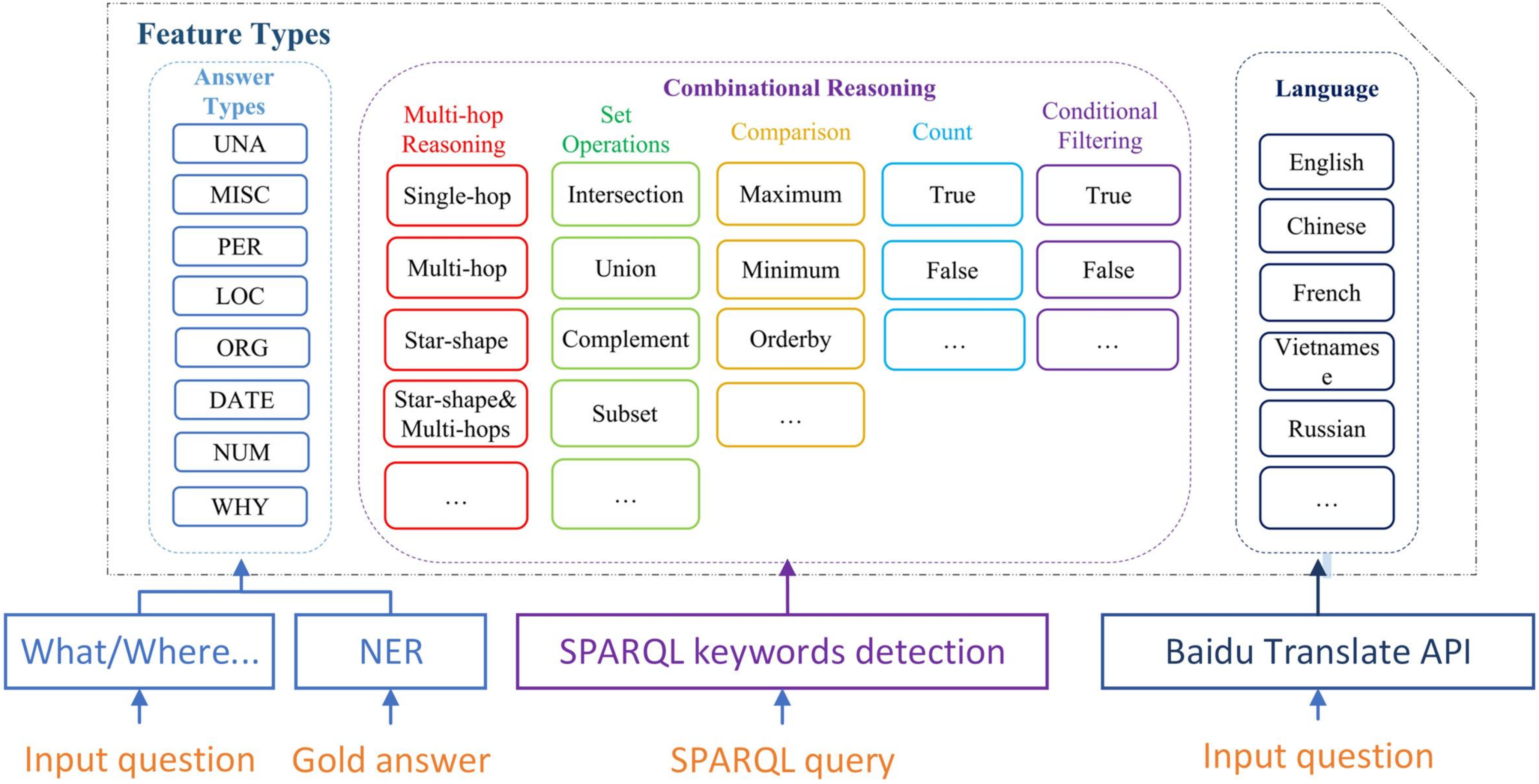
Reasoning Type: From inference type labels in existing KBQA datasets and keywords involved in SPARQL queries.

Language Type: From language labels in existing multilingual KBQA datasets.

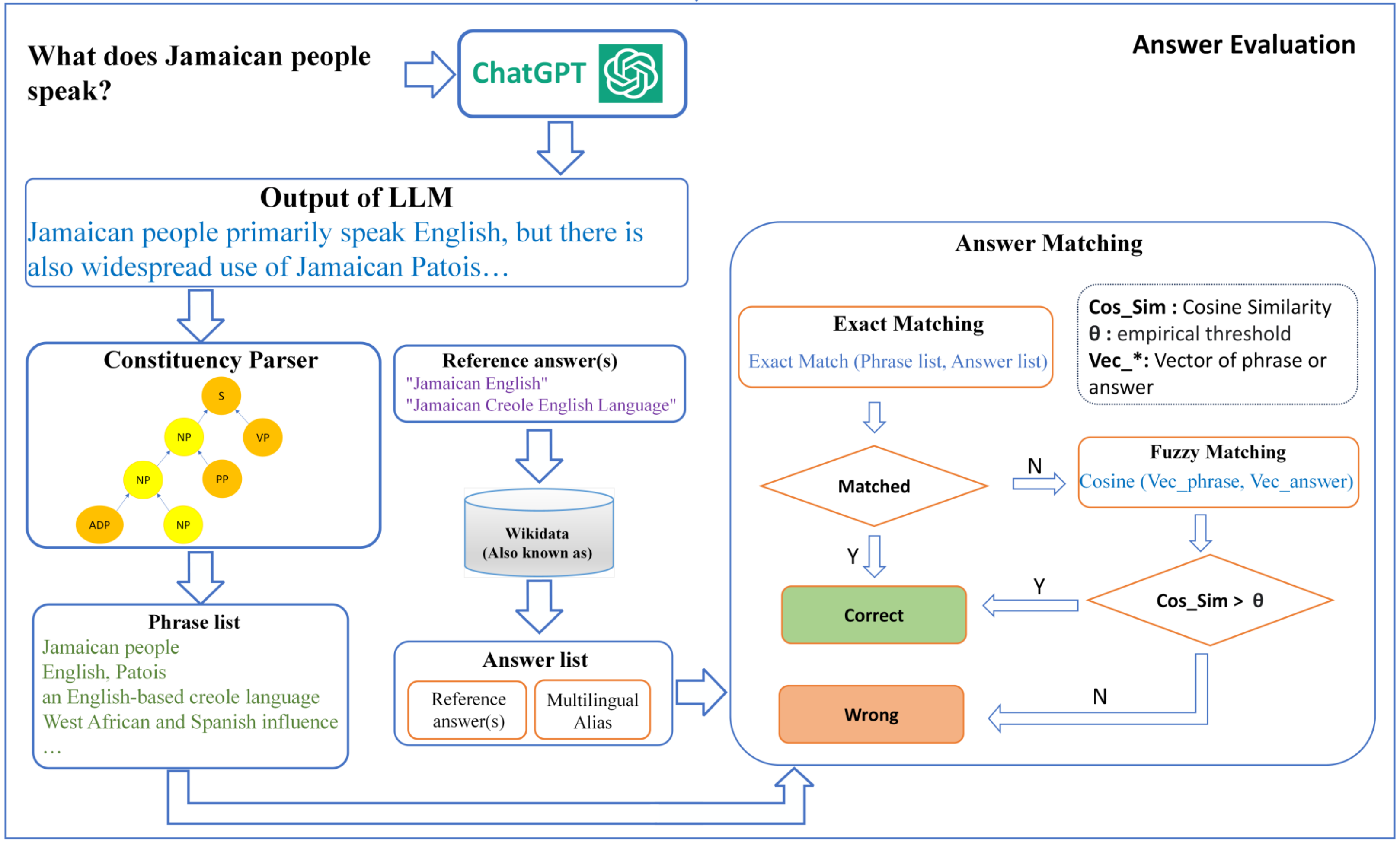
Table 1. The feature-driven question tags defined in this paper.

Answer type	Description
MISC	The answer to the question is the miscellaneous fact defined by the named entity recognition task.
PER	The answer to the question is the name of a person.
LOC	The answer to the question is a location.
WHY	The answer explains the reasons for the facts mentioned in the question.
DATE	The answer to the question is a date or time.
NUM	The answer to the question is a number.
Boolean	The answer to the question is yes or no.
ORG	The answer to the question is the name of a organization.
UNA	The input question is unable to answer.
Reasoning type	**Description**
SetOperation	The process of obtaining answers involves set operations.
Filter	The answer is obtained through condition filtering.
Counting	The process of obtaining an answer involves counting operations.
Comparative	The answer needs to be obtained by comparing or sorting numerical values.
Single-hop	Answering questions requires a single-hop Reasoning.
Multi-hop	Answering questions requires multi-hop Reasoning.
Star-shape	The reasoning graph corresponding to inputting question is star-shape.

Question Labeling – Auto labeling



Answer Evaluation – Answer Matching



Expanded Exact Matching:

We obtained multilingual aliases for all reference answers from Wikipedia, greatly expanding the matching scope of the Gold list.

Fuzzy Matching:

Fuzzy matching is performed based on cosine similarity thresholds using m-BERT word vectors.

Condition for fuzzy matching :

when EM fails and the answer type is not a number, date, symbol code, or other sequences that are difficult to distinguish based on vector similarity.

Answer Evaluation – CheckList testing

Original Test case:
What unit does the international system of units use to measure magnetic flux density?

INV cases

Case 1: Provide some orthographic variations (potentially erroneous)

What unit does the international system of units use to measure magneti flux density?

Case 2: Paraphrase

What unit does the international system of units use to measure magnetic flux density?

↓

Which unit is utilized by the International System of Units for measuring magnetic flux density?

→

INV Metric

Record positive instances when the model produces the same judge result for the output of the three inputs.

DIR cases

Case 1: Altering the execution details of reasoning

What unit does the international system of units do not use to measure magnetic flux density? Generate the corresponding SPARQL query.

Case 2: Add prompt with answer type info

What unit does the international system of units use to measure magnetic flux density?, the type of answer is 'miscellaneous'.

Case 3: CoT (step-by-step):

Input1: What does 'unit' mean?
Input2: What does 'international system' mean?
Input3: What does 'measure magnetic' mean?
Input4: …
Input5: What unit does the international system of units use to measure magnetic flux density?

DIR Metric

When the model output matches the expected output of the cases, it is recorded as a positive instance.

Case 1 expect [correct revise in SPARQL]:
SPARQL with a new filter process.

Case 2 expect [matched answer type]:
The type of answer generated by the model matches/corresponds to the answer type provided in the prompt.

Case 3 expect [improved accuracy of answers]:
Generating answers with higher accuracy.

Part 4

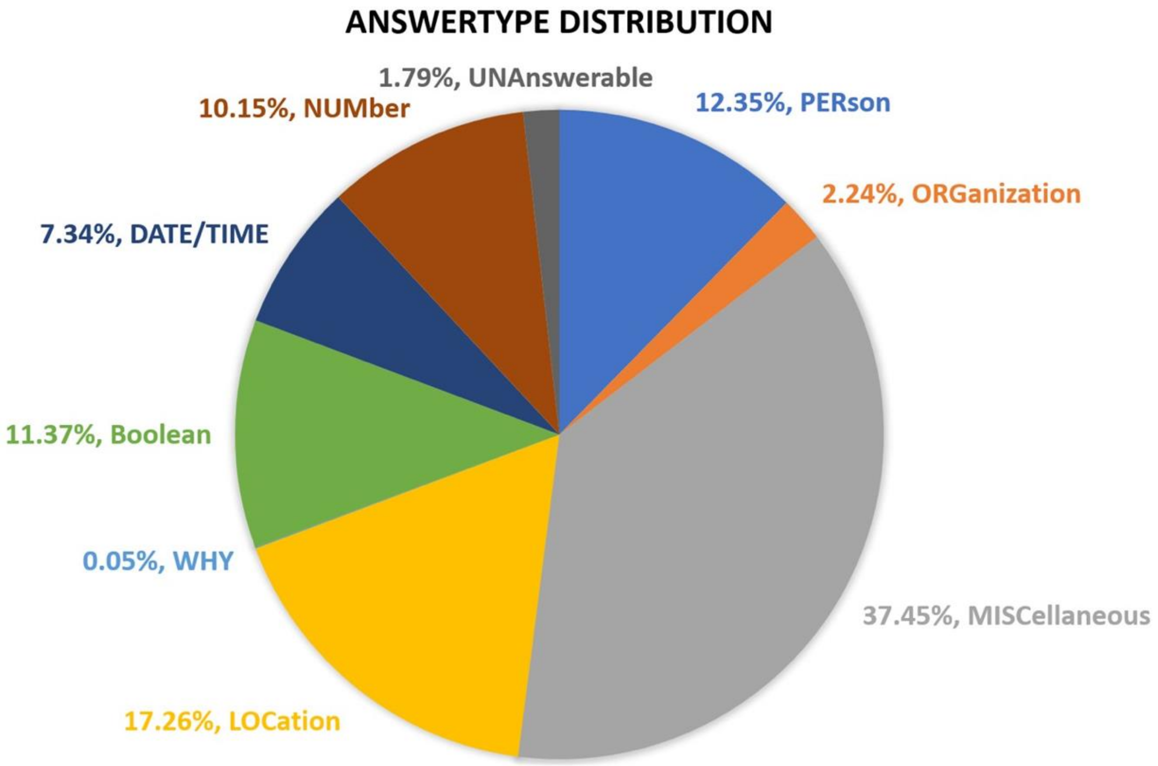
Experiments and key findings

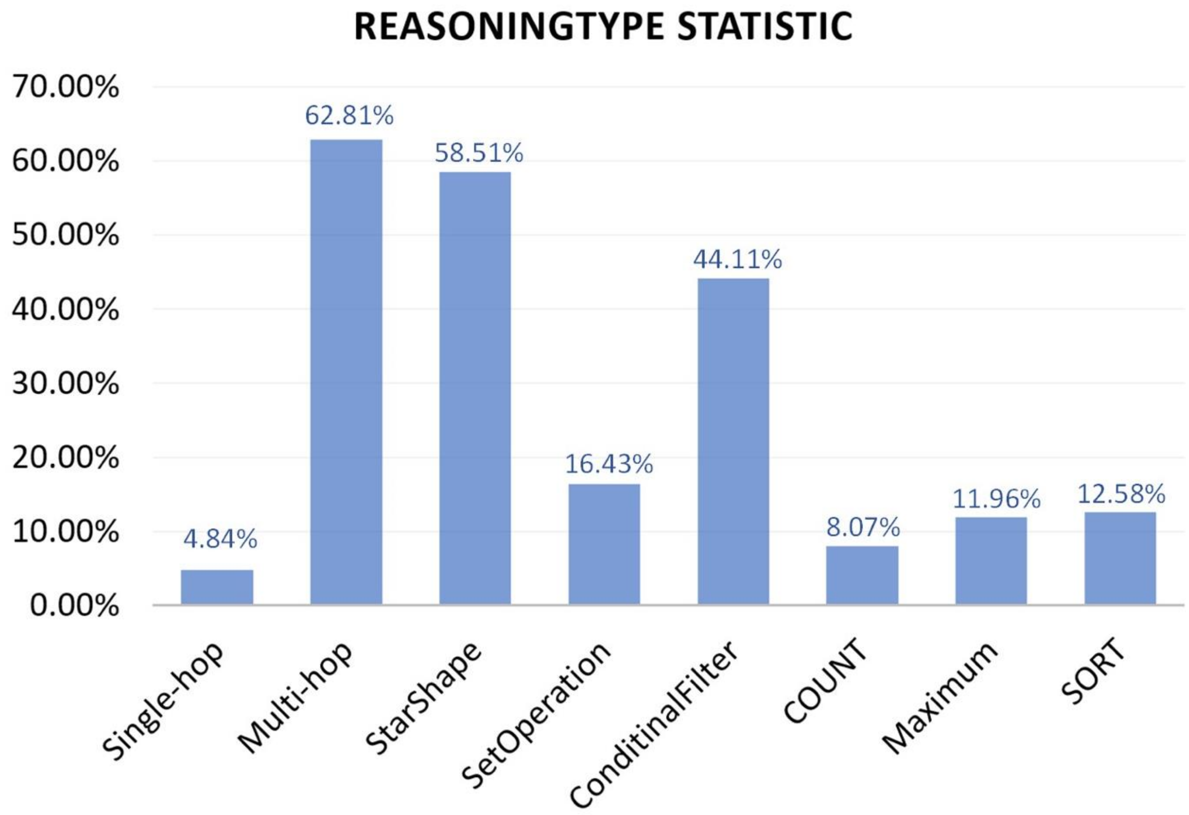
Results and key findings – Datasets for testing

Table 2. The Statistical of collected KB-based CQA datasets, "Col. Size" represents the size of the dataset we collected in our experiments. "Size" denotes the original size of the dataset.

	Datasets	Size	Col. Size	Lang
(Cao, S. et. al, 2022)	KQApro	117,970	106,173	EN
(Dubey, M. et. al, 2019)	LC-quad2.0	26,975	26,975	EN
(Yih, W.t. et. al, 2016)	WQSP	4737	4,700	EN
(Talmor, A. et. al, 2018)	CWQ	31,158	31,158	EN
(Gu, Y. et. al, 2021)	GrailQA	64,331	6,763	EN
(Su, Y. et. al, 2016)	GraphQ	4,776	4,776	EN
(Ngomo, N. et. al, 2018)	QALD-9	6,045	6,045	Mul
(Longpre, S. et. al, 2021)	MKQA	260,000	6,144	Mul
	Total Collected		194,782	

Results and key findings - Datasets for testing





GPT family:

GPT-3 (text-davinci-001)

GPT-3.5 v2 (text-davinci-002)

GPT-3.5 v3 (text-davinci-003)

ChatGPT (gpt3.5-turbo-0301)

GPT-4

LLM not belongs to GPT family :

FLAN-T5 (Text-to-Text Transfer Transformer 11B)

Main results – Overall Testing

Table 3. Overall results of the evaluation. We compare the exact match of ChatGPT with current SOTA traditional KBQA models (fine-tuned (FT) and zero-shot (ZS)), GPT family LLMs, and Non-GPT LLM. In GraphQ, QALD-9 and LC-quad2, the evaluation metric used is F1, while other datasets use Accuracy (Exact match).

Datasets	KQApro	LC-quad2	WQSP	CWQ	GrailQA	GraphQ	QALD-9	MKQA
	Acc	F1	Acc	Acc	Acc	F1	F1	Acc
SOTA(FT)	**93.85** [29]	33.10 [31]	73.10 [15]	**72.20** [15]	**76.31** ‡	31.8 [13]	**67.82** [32]	46.00 [22]
SOTA(ZS)	94.20 [25]	-	62.98 [50]	-	-	-	-	-
FLAN-T5	37.27	30.14	59.87	46.69	29.02	32.27	30.17	20.17
GPT-3	38.28	33.04	67.68	51.77	27.58	38.32	38.54	26.97
GPT-3.5v2	38.01	33.77	72.34	53.96	30.50	40.85	44.96	30.14
GPT-3.5v3	40.35	39.04	79.60	57.54	35.43	47.95	46.19	39.05
ChatGPT	47.93	42.76	83.70	64.02	46.77	53.10	45.71	44.30
GPT-4	57.20	**54.95**	**90.45**	71.00	51.40	**63.20**	57.20	**59.20**

Main results – Overall Testing

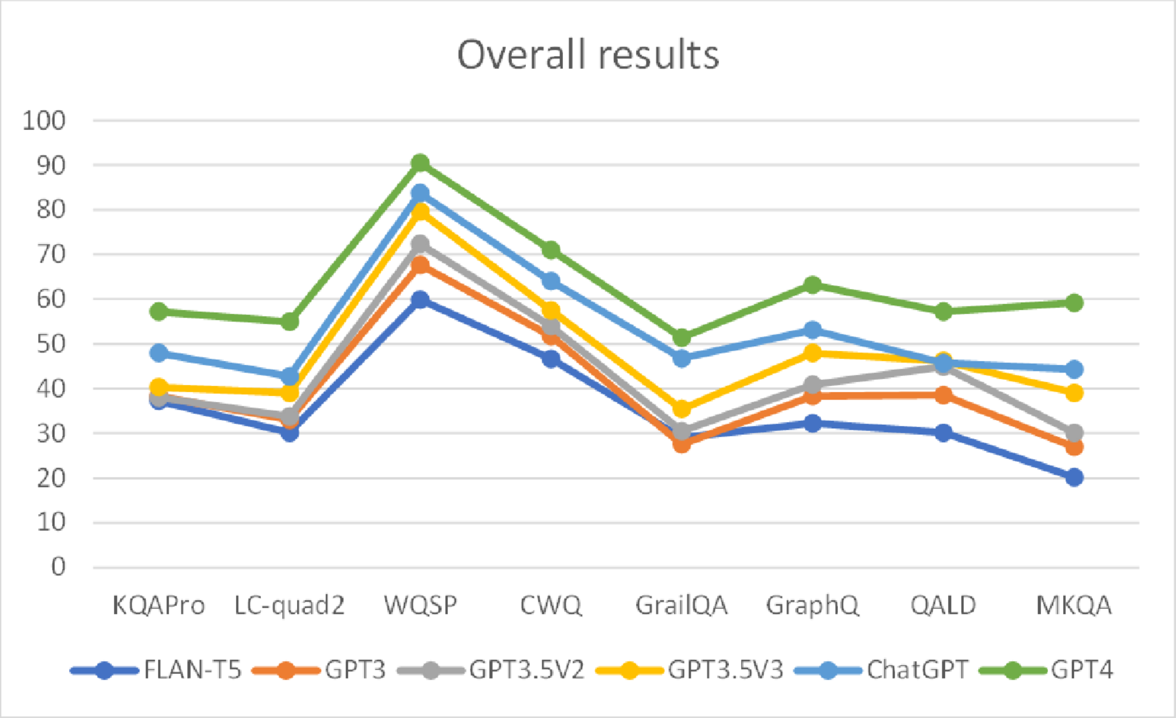
Table 4. Comparison of LLMs on multilingual test sets.

Languages	FLAN-T5	GPT-3	GPT-3.5v2	GPT-3.5v3	ChatGPT	GPT-4
en	30.29	57.53	56.99	64.16	**66.49**	66.09
nl	20.75	50.47	54.58	60.56	65.05	**69.72**
de	22.40	50.54	54.48	57.17	62.54	**73.91**
es	21.68	48.22	55.70	58.50	**61.87**	57.69
fr	26.16	49.46	55.02	57.89	**62.19**	62.00
it	24.19	47.67	52.33	58.06	58.96	**73.91**
ro	22.28	44.38	50.94	54.12	59.55	**63.41**
pt_br	15.38	38.46	38.46	42.31	50.00	**66.67**
pt	20.58	37.70	44.26	50.27	**52.64**	52.25
ru	7.29	20.58	29.69	21.68	32.24	**49.58**
hi_in	3.61	9.93	19.13	13.54	21.48	**25.00**
fa	2.45	6.59	21.09	11.49	22.03	**31.71**
zh_cn	3.65	17.45	22.40	24.87	33.46	**44.62**

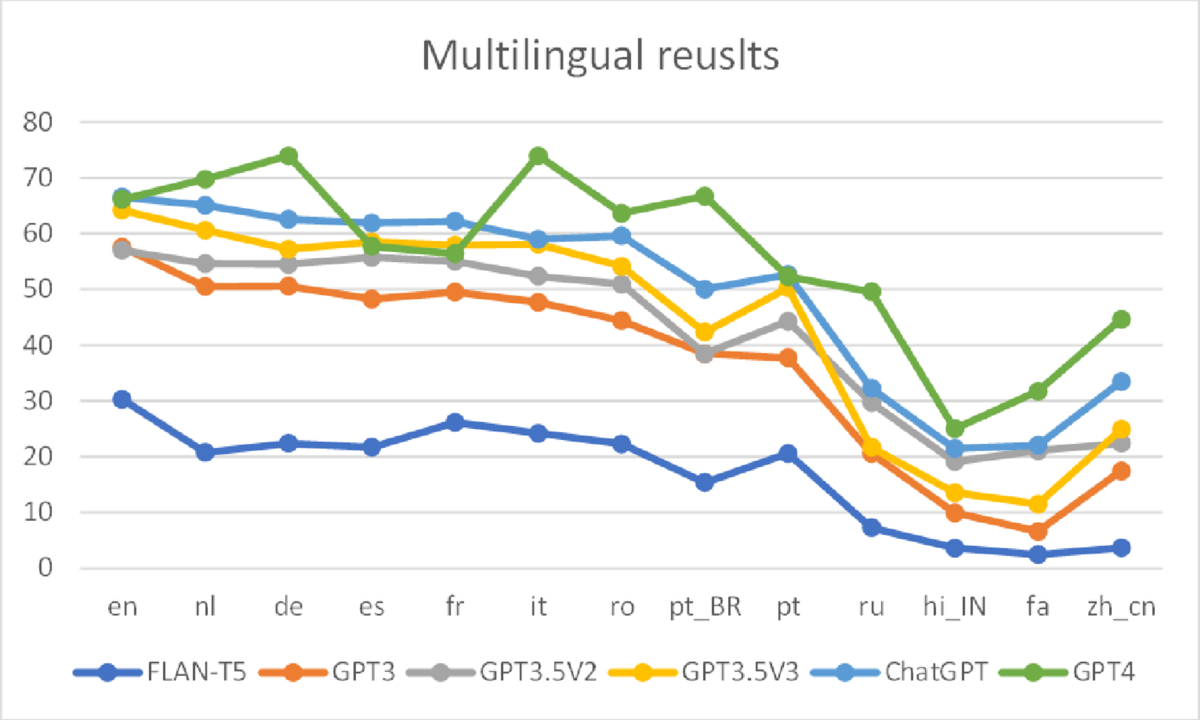
1. With each new iteration, the GPT family's multilingual question-answering capabilities are on the rise.

2. The improvement of GPT-4 indicates that the introduction of multimodal information significantly enhances performance for certain language types

Main results – Overall Testing



(a)



(b)

From a dataset perspective, the GPT models and FLAN-T5 share a high degree of similarity in their trendlines.

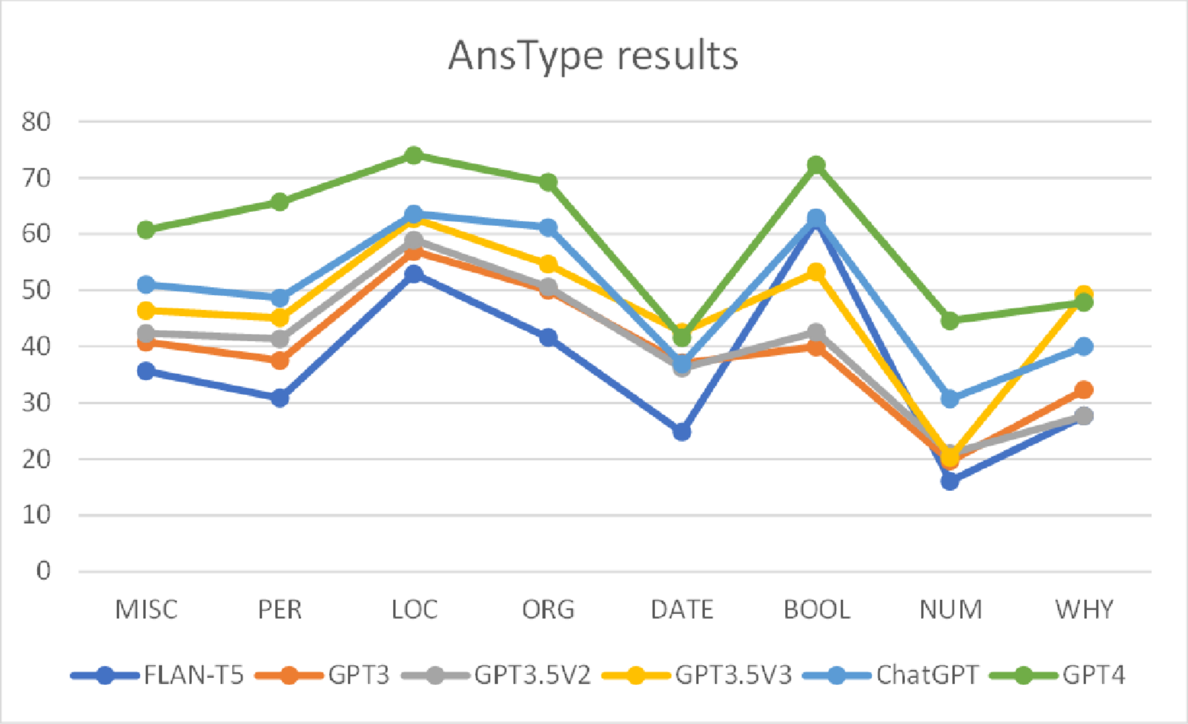
From a multilingual question-answering perspective, before the introduction of multimodal information (GPT-4), the GPT family also maintained a roughly similar trendline shape.

Main results – Overall Testing

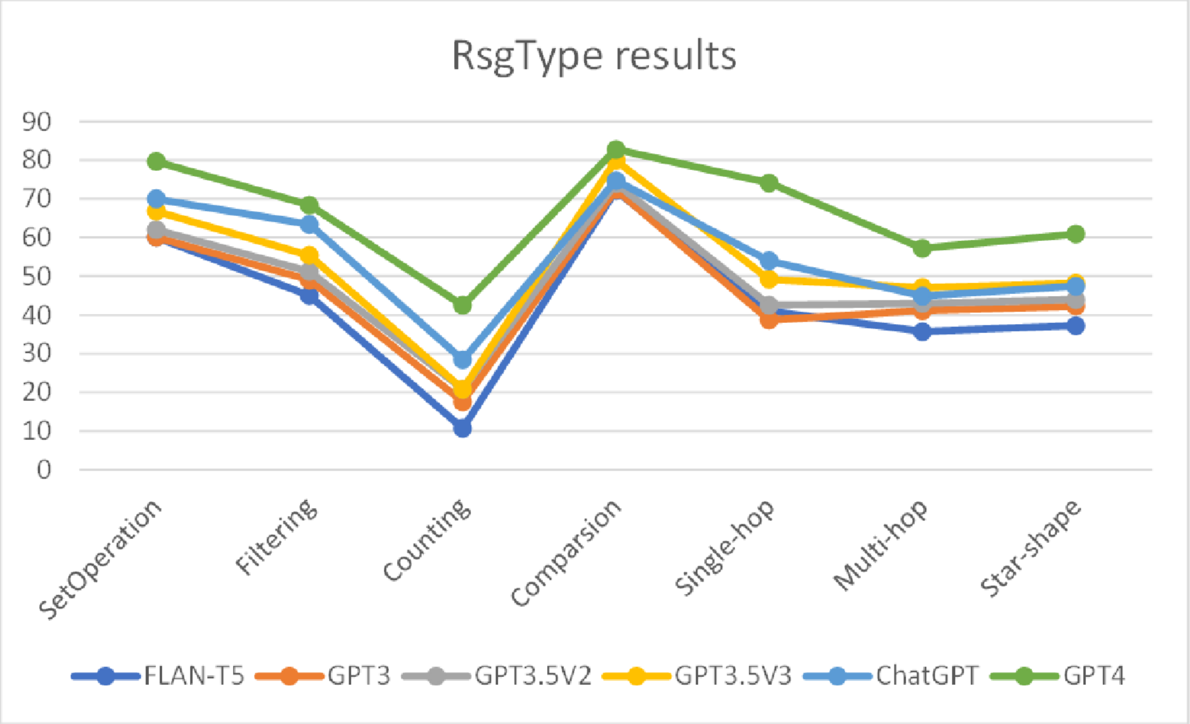
Table 5. Exact Match comparison based on Answer Types (AnsType) and Reasoning Types (RsgType)

MF	FLAN-T5	GPT-3	GPT-3.5v2	GPT-3.5v3	ChatGPT	GPT-4
AnsType						
MISC	35.67	40.79	42.35	46.42	51.02	**60.73**
PER	30.84	37.53	41.36	45.10	48.65	**65.71**
LOC	52.91	56.92	58.93	62.71	63.55	**73.98**
ORG	41.62	50.01	50.58	54.62	61.18	**69.20**
DATE	24.81	37.07	36.15	42.54	36.92	**41.57**
Boolean	62.43	39.96	42.56	53.23	62.92	**72.28**
NUM	16.08	19.66	21.01	20.31	30.70	**44.59**
WHY	27.69	32.31	27.69	**49.23**	40.00	47.83
UNA	-	-	-	-	-	-
RsgType						
SetOperation	60.11	60.12	62.03	66.86	70.00	**79.70**
Filtering	45.01	49.06	51.24	55.43	63.40	**68.40**
Counting	10.68	17.56	20.83	20.83	28.41	**42.50**
Comparison	72.13	72.44	74.00	80.00	74.74	**82.79**
Single-hop	41.00	38.72	42.54	49.22	54.00	**74.14**
Multi-hop	35.68	41.09	42.98	47.06	44.88	**57.20**
Star-shape	37.23	42.28	43.96	48.17	47.43	**60.91**

Main results – Overall Testing



(c)



(d)

In terms of the types of answers to questions, there's a striking similarity in the strengths and weaknesses of past GPT models and FLAN-T5.

In terms of the types of reasoning involved in the questions, FLAN-T5 and the GPT family tend to excel or struggle with the same kinds of reasoning operations.

CheckList results

Table 6. MFT results of ChatGPT

	SetOperation	Filtering	Counting	Comparison	Single-hop	Multi-hop	Star-shape
Single Reasoning	60.22	51.39	24.16	31.48	44.07	**48.27**	**50.75**
Multiple Reasoning	**70.00**	**63.40**	**28.41**	**74.74**	**54.00**	44.88	47.43

Table 7. INV results of GPT family

LLM	CCC	CCW	CWC	CWW	WCC	WCW	WWC	WWW	Stability Rate
GPT-3	434	64	59	52	42	43	73	666	76.76
GPT-3.5 v2	495	44	65	42	43	30	58	656	80.30
GPT-3.5 v3	604	46	43	49	34	35	49	583	82.83
ChatGPT	588	49	72	68	52	27	32	545	79.06
GPT-4	798	0	0	65	54	0	0	516	**91.70**

Table 8. DIR results for RsgType, the score represents the percentage of expected output produced by the LLMs.

	SetOperation	Filtering	Counting	Comparison	Overall
GPT-3.5 v3	45%	75%	65%	**65%**	62.5%
ChatGPT	**75%**	85%	**70%**	**65%**	**73.75%**
GPT-4	65%	**90%**	**70%**	60%	71.25%

MFT result

Multiple types of reasoning better than single type of reasoning

INV result

The consistency of the GPT model has steadily improved with each iteration, approaching the trend of traditional models.

DIR case 1

ChatGPT produce responses that aligned more closely with expectations for the DIR test case 1

CheckList results

Table 9. DIR results for AnsType prompting

	MISC	PER	LOC	ORG	DATE	Boolean	NUM	WHY
GPT-3	+1.43	0	+5.71	+4.29	+4.29	+15.71	+17.14	0
GPT-3.5 v2	-4.28	+2.85	+7.14	+14.28	+2.86	-8.57	+14.28	+12.13
GPT-3.5 v3	-12.86	+10.00	+18.57	-7.14	+4.71	+17.14	+22.85	+9.09
ChatGPT	+6.78	-3.64	-1.72	-5.35	-8.58	+4.28	+7.15	-3.03
GPT-4	-4.29	-2.86	+11.43	+5.71	0	+7.14	+4.29	-6.06

DIR case 2

Answer type prompting produces better results for weaker models.

Table 10. DIR results for CoT prompting

	MISC	PER	LOC	ORG	DATE	Boolean	NUM	WHY
GPT-3	-1.40	-2.00	-2.67	+2.73	-3.77	+3.36	+35.66	+6.06
GPT-3.5 v2	-0.35	-5.33	+1.78	-3.64	+0.76	-5.04	+32.95	0
GPT-3.5 v3	0	-2.00	-1.33	-1.82	-1.51	-2.10	+34.12	0
ChatGPT	-1.75	-4.66	+0.89	-3.63	-1.50	+3.36	+30.62	+6.06
GPT-4	-3.00	+11.11	+2.22	+3.3	-2.71	0	+20.00	+2.62
	SetOperation	Filtering	Counting	Comparison	Multi-hop	Star-shape		
GPT-3	+10.79	+10.43	+35.66	+1.35	-1.60	-1.69		
GPT-3.5 v2	+4.86	+5.46	+38.54	-2.26	-1.18	-0.85		
GPT-3.5 v3	+6.34	+8.18	+38.99	-1.13	-1.61	-1.26		
ChatGPT	+7.82	+9.47	+35.78	+0.45	-1.47	-1.41		
GPT-4	+2.05	+0.93	+11.11	-1.88	+2.82	+2.68		

DIR case 3

Multi-step prompting can significantly enhance LLM's ability to tackle specific types of questions.

Part 5

Conclusion

Q1: Can LLM replace traditional KB or become a new form of KB?

A1: The precondition is that we need to find LLM-specific SPARQL so that it can access the knowledge it contains correctly and reliably.

Q2: Can GPT models based on their own knowledge potentially replace traditional KBQA models?

A2: Not yet, although on some test sets, GPT-4's QA performance has exceeded traditional models. However, its lower consistency makes it not a reliable QA model.

Thank you !