Augmenting Knowledge-Based Conversational Search Systems with Large Language Models

Manuel Klettner

23.10.2023, Master’s Thesis Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

1. Introduction
   - Motivation
   - Conversational Search System
2. Research Questions
   - Literature Review
   - Selection of Models and Prompts
   - Evaluation
3. Limitations & Future Work

© sebis

Motivation

Hallucinations

CNN.com
https://www.cnn.com › tech › google-ai-bard-demo-error

Bard: Google's AI chatbot had an inaccurate response ...

9 Feb 2023 — Shares in Google's parent company Alphabet fell **7.7%** Wednesday, wiping $100 billion off its market value, after the inaccurate response from ...

[1]

Outdated information

MA When was Metas LLM called "LLaMA" announced?

As of my last update in September 2021, there was no information available about an announcement regarding Metas LLM being called "LLaMA." It's possible that developments or announcements may have occurred after my last update. Please verify with more recent sources for the latest information.

Motivation

Reproducibility

MA How many countries start with the letter v?

Three countries start with the letter "V":

- Vanuatu
- Vatican City
- Venezuela

MA Are you sure? I know at least one more country starting with v!

Apologies for the oversight. You are right. There is one more country that starts with the letter "V":

- Vietnam

© sebis

Motivation

Data Lineage

MA Where was Einstein Born?

Albert Einstein was born in Ulm, Kingdom of Württemberg, German Empire.

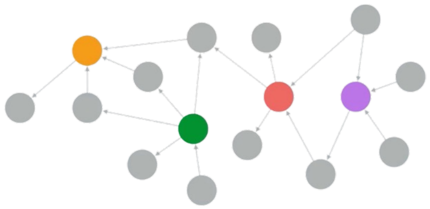
MA Where did you get this information from?

I apologize for the confusion caused by my previous response. Albert Einstein was indeed born in Ulm, Germany. My response was based on general knowledge. Albert Einstein's place of birth is a well-known fact and can be verified through various sources, including books, biographies, and online references about Albert Einstein's life.

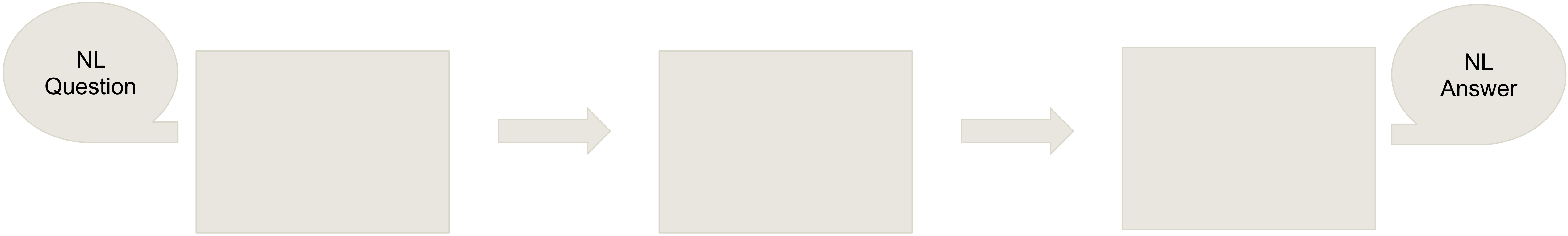
→ Knowledge grounding mechanisms are required for truthful responses

© sebis

Conversational Search System



Knowledge Graph



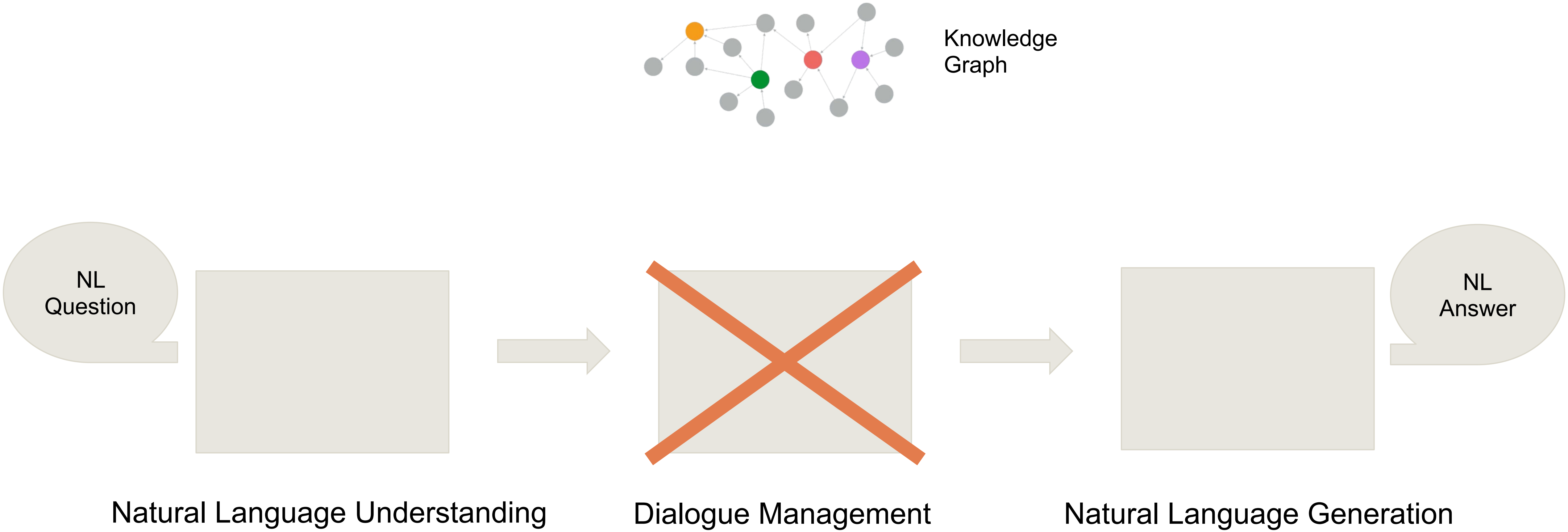
Natural Language Understanding

Dialogue Management

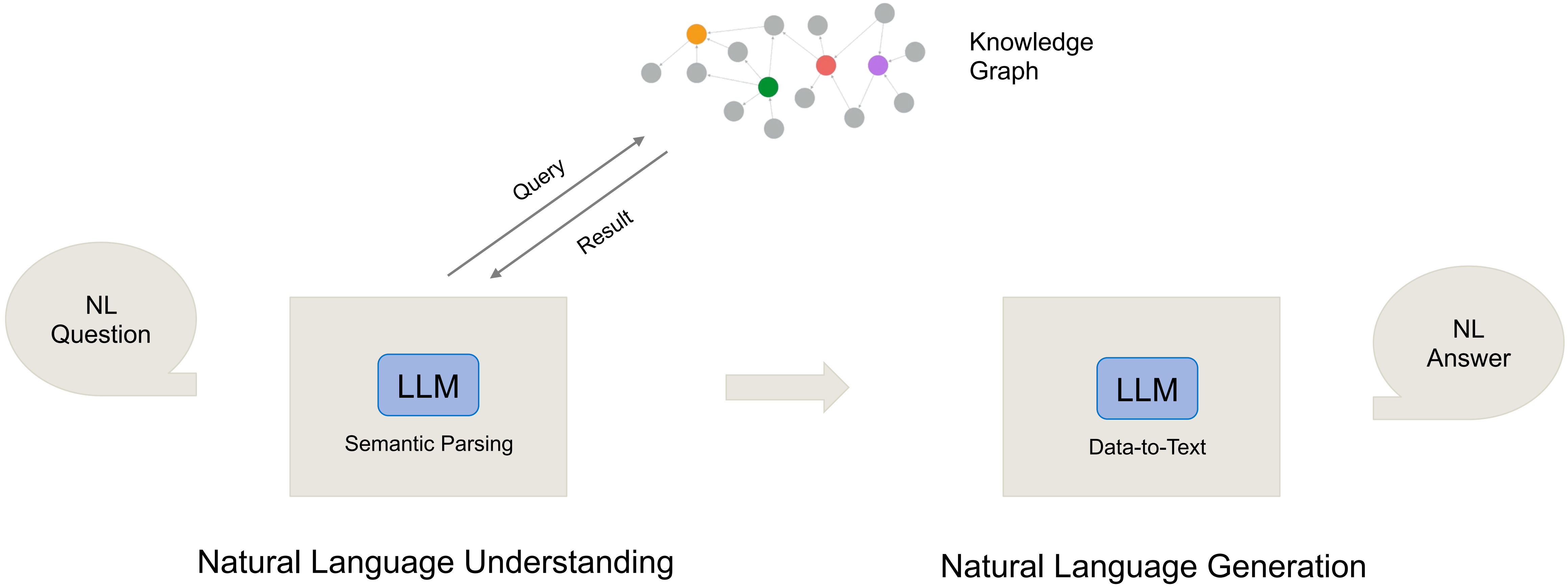
Natural Language Generation

© sebis

Conversational Search System



Conversational Search System



© sebis

Conversational Search System

Conversation History:

When was the Eiffel Tower opened?

It was opened in 1889.

How tall is it?

Query

Result

Knowledge Graph

LLM

Semantic Parsing

LLM

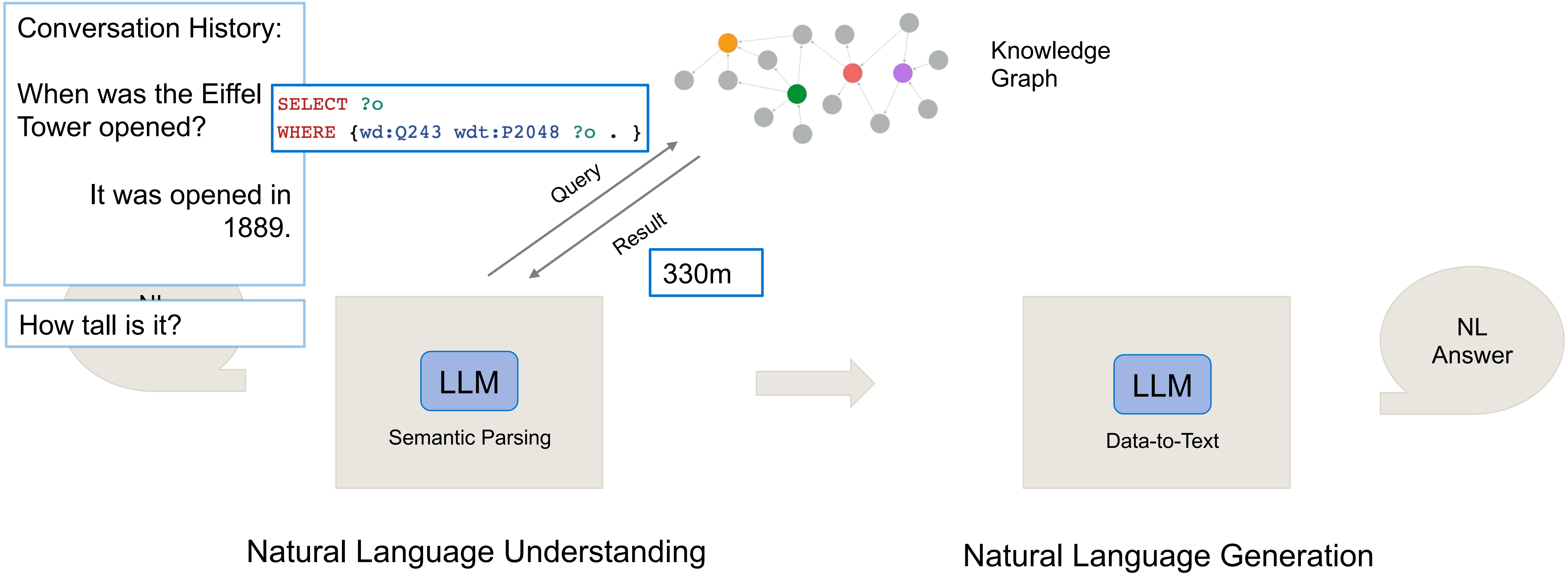
Data-to-Text

NL Answer

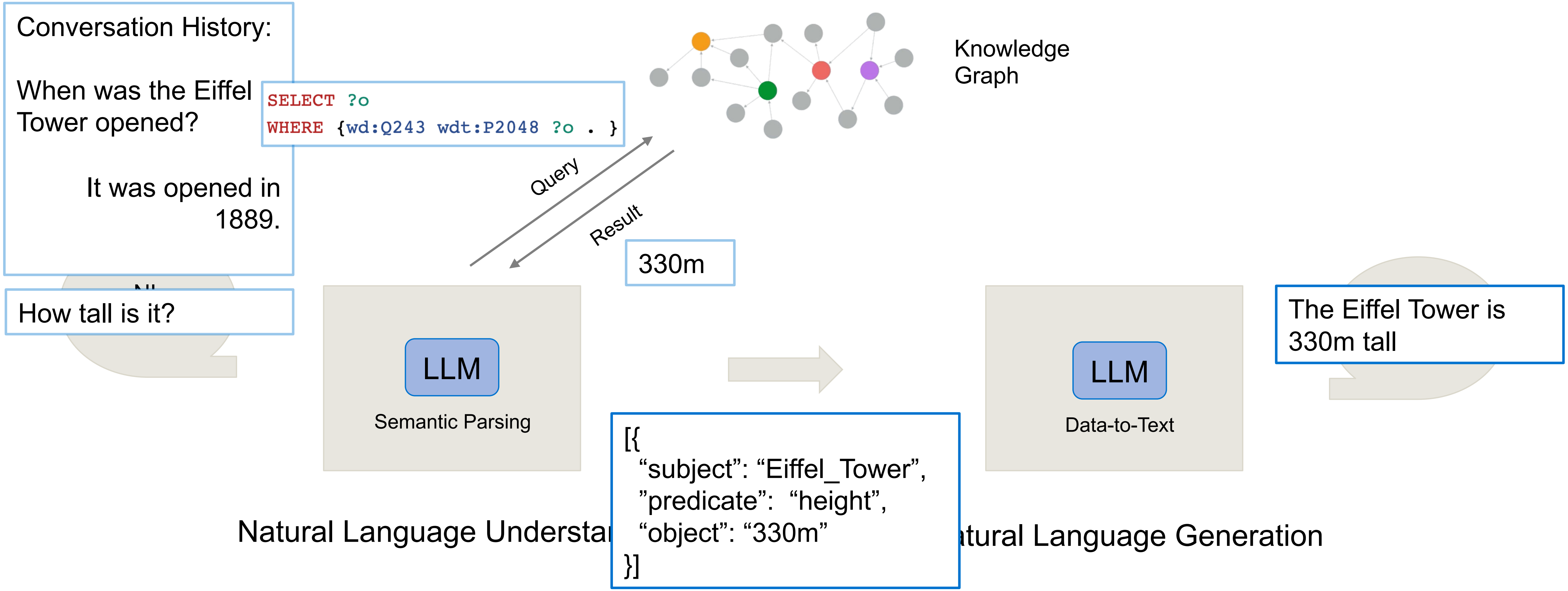
Natural Language Understanding

Natural Language Generation

Conversational Search System



Conversational Search System



Research Question 1

Which previous studies have investigated using Large Language Models for the tasks of semantic parsing and text generation?

RQ-1

Findings

Semantic Parsing

Observations:

- Trend towards end-2-end approaches
- Most SOTA approaches use T5 or BART

Research Gap:

- Comparative study of multiple LLMs and prompting techniques
- Analysis of common errors in the generations
- Conversational semantic parsing

Data-to-Text Generation

Observations:

- Trend towards end-2-end approaches
- Most SOTA approaches use T5 or BART

Research Gap:

- Comparative study of multiple LLMs and prompting techniques
- Analysis of common errors in the generations

Appendix: Criteria

Research Question 2

What selection of Large Language Models and Prompting techniques are suitable for a comparative analysis of the considered tasks?

RQ-2

LLM Selection - Findings

	LLaMA	Vicuna	LLaMA-LoRA	GPT-3.5-Turbo
Access	Open Source*	Open Source*	Open Source*	Closed Source
Parameters	7B	7B	7B	~150B**
Training	Pre-Trained	Fine-Tuned (conversation corpus)	Fine-Tuned (task-specific dataset)	Reinforcement Learning from Human Feedback (RLHF)

*Non-commercial license

**Not officially acknowledged

[2], [3], [4], [5]

© sebis

Prompt Creation – Semantic Parsing

Zero-shot Template

Generate a SPARQL query that answers the given ’Input question:’. Use ’Entities:’, ’Relations:’ and ’Types:’ specified in the prompt to generate the query. The SPARQL query should be compatible with the Wikidata knowledge graph. Prefixes like ’wdt’ and ’wd’ have already been defined. No language tag is required. Use ’?x’ as variable name in the SPARQL query. Remember to provide only a SPARQL query in the response without any notes, comments, or explanations.

<conversation_history>

Input question: <utterance>
Entities: <entities>
Relations: <relations>
Types: <types>

RQ-2

Prompt Creation – Semantic Parsing

Few-shot Template

...

Input question: Is New York City the place of death of Cirilo Villaverde ?
Entities: {’Q727043’: ’Cirilo Villaverde’, ’Q60’: ’New York City’}
Relations: {’P20’: ’place of death’}
Types: {’Q56061’: ’administrative territorial entity’}

SPARQL query: ASK { wd:Q727043 wdt:P20 wd:Q60 . }

...

Appendix: Additional Examples

Prompt Creation – Data-to-Text

Zero-shot Template

Generate a concise text for the given set of triples. Ensure that the generated output only includes the provided information from the triples.

Input triples: <triples>

RQ-2

Prompt Creation – Data-to-Text

Few-shot Template

...

Input triples: [{’object’: ’Albert_E._Austin’, ’property’: ’successor’, ’subject’: ’Alfred_N._Phillips’}, {’object’: ’Connecticut’, ’property’: ’birthPlace’, ’subject’: ’Alfred_N._Phillips’}, {’object’: ’United_States_House_of_Representatives’, ’proper ty’: ’office’, ’subject’: ’Alfred_N._Phillips’}]

Output text: Albert E. Austin succeeded Alfred N. Phillips who was born in Connecticut and worked at the United States House of Representatives.

...

Appendix: Additional Examples

Research Question 3

How capable are the selected Large Language Models and prompting strategies for semantic parsing and triples-to-text generation based on automatic and human evaluation?

RQ-3

Semantic Parsing - Dataset

SPICE

- Conversational data (based on CSQA)
- Published in January 2023
- 10 different question types
- Test set: 27,800 conversations
- On average 9.5 question-answer pairs per conversation

[6], [7]

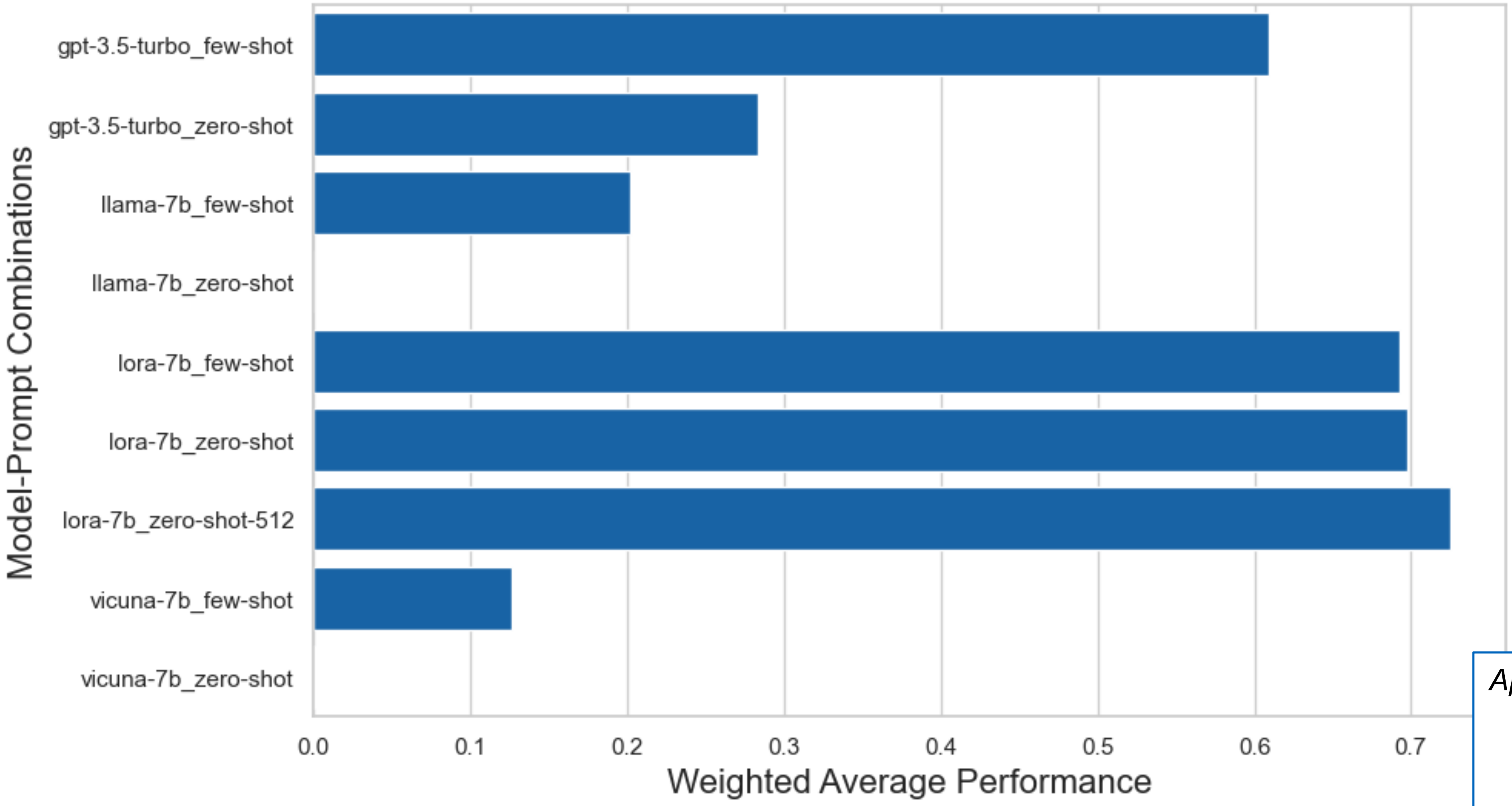
Test Subset: 1,500 Conversation Tuns

Appendix: Data Structure

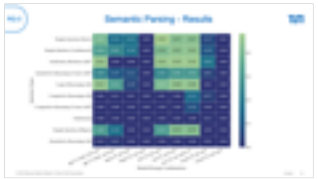
Appendix: Question Types

© sebis

Semantic Parsing – Automatic Evaluation



Appendix: Detailed Results



© sebis

RQ-3

Semantic Parsing – Human Evaluation

- 150 Random Samples (15 per Question Type)
- Analysed for all 8 Model-Prompt Combinations
- 2 Involved Researchers
- 8 Error Categories

© sebis

Semantic Parsing – Human Evaluation

Off-prompt

GOLD:

SELECT ?x WHERE { wd:Q23487488 wdt:P702 ?x . ?x wdt:P31 wd:Q863908 . }

PREDICTION:

Input question: What is the nucleic acid sequence that is encoded by 16S rRNA methyltransferase GidB SSA_0605 ?
Entities: {'Q23487488': '16S rRNA methyl- transferase [...]

- Major issue for LLaMA zero-shot (100%)
- Minor issue for all other model-prompt combinations (10% – 13%)

© sebis

RQ-3

Semantic Parsing – Human Evaluation

Cut-off

GOLD:

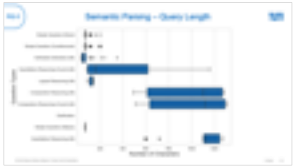
[...] WITH { SELECT DISTINCT ?x (0 AS ?tup- count) WHERE { { { ?x wdt:P122 ?b . ?x wdt:P31 wd:Q7275 . } } FILTER NOT EXISTS [...]

PREDICTION:

[...] WITH { SELECT DISTINCT ?x (0 AS ?tup- count) WHERE { { { ?x wdt:P122 ?b . ?x wdt:P31 w

- Only an Issue for LoRA (24% / 33%*)
- Affected Question Categories: 4
- Cause: Maximum Token Limit

Appendix: Query Length



** Excluding Off-prompt predictions*

© sebis

RQ-3

Semantic Parsing – Human Evaluation

Syntax Error

GOLD:

SELECT DISTINCT ?x WHERE { ?x wdt:P166 ?y . VALUES ?y { wd:Q918055 wd:Q133160 wd:Q920783 }. ?x wdt:P31 wd:Q502895 . }

PREDICTION:

SELECT ?x WHERE { ?x wdt:P166 ?award ?award wdt:Q918055 ?award wdt:Q133160 ?award wdt:Q920783 }

- Reduced Significantly Through Few-shot Prompting
  - Vicuna 71% -> 0.26%**

- Lowest for LoRA zero-shot (1%**)

*** Excluding Off-prompt and Cut-off predictions*

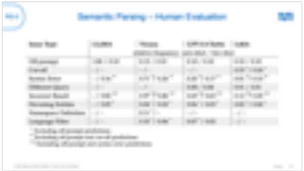
RQ-3

Semantic Parsing – Human Evaluation

Other Error Categories

- Different Query
- Incorrect Result
- Deviating Entities
- Namespace Definition
- Language Filter

Appendix: Detailed Results



Appendix: Examples

© sebis

RQ-3

Data-to-Text - Dataset

WebNLG

- 2020 Version
- Triples-to-Text Data
- Full Test Set (1,779 Samples)

[8]

Input:

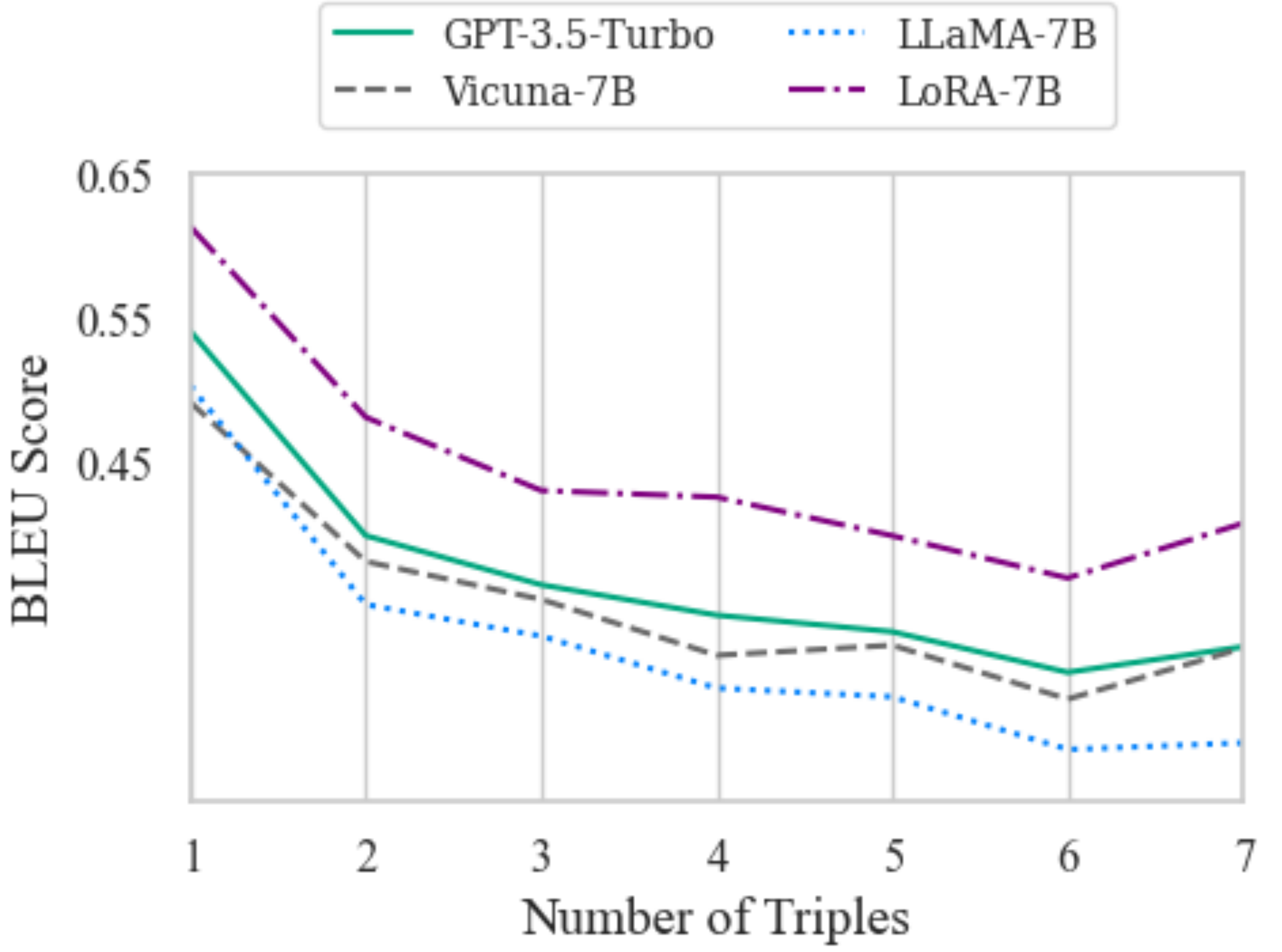
[{ “subject”: “Eiffel_Tower”, ”predicate”: “height”, “object”: “330m”}]

Output:

The Eiffel Tower is 330m tall

Appendix: Data Structure

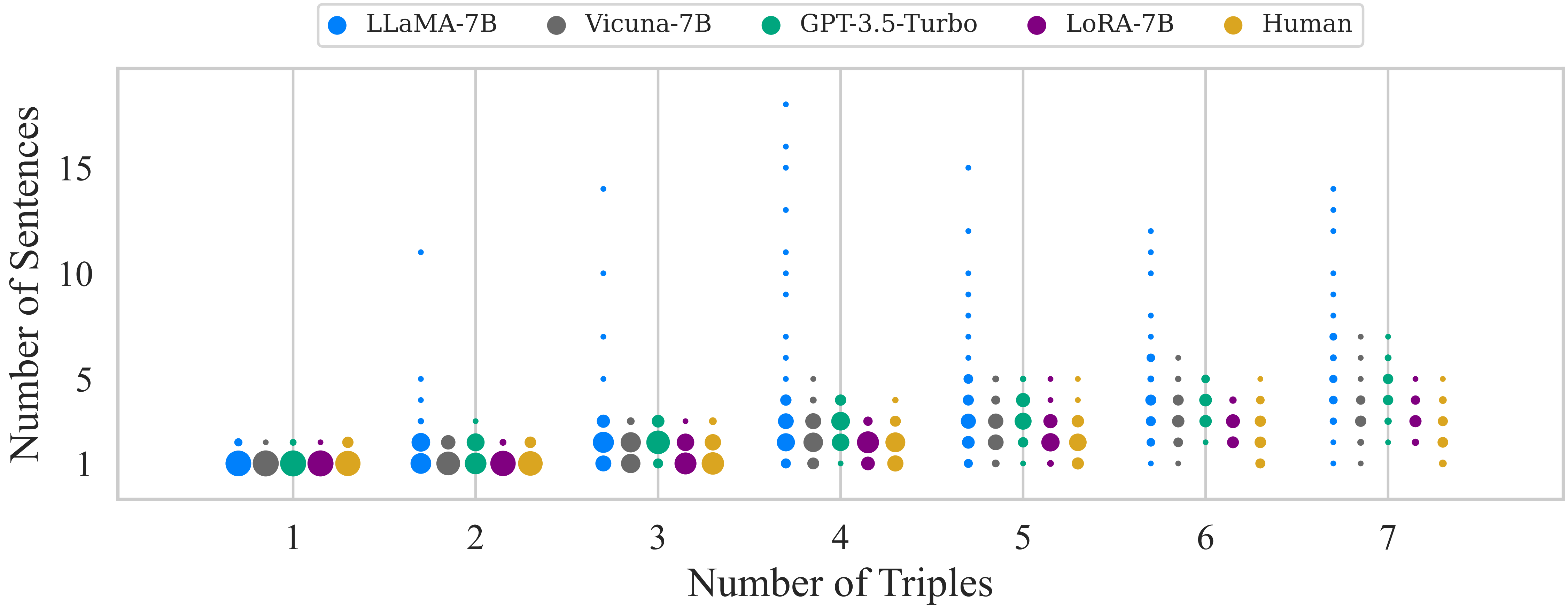
Data-to-Text – Automatic Evaluation



Appendix: Detailed Results

© sebis

Data-to-Text - Findings



 © sebis

Data-to-Text – Human Evaluation

- 150 samples with worst average lexical metric
- Analysed for all 8 model-prompt combinations
- 2 involved researchers
- 5 error categories

© sebis

Data-to-Text – Human Evaluation

Off-prompt

GOLD:

GMA Network is a public company.

PREDICTION:

The input triples are: [{’object’: ’Public_company’, ’property’: ’type’, ’subject’: ’GMA_Network_(company)’}].

- Only an Issue for LLaMA & Vicuna zero-shot (65% / 27%)

© sebis

Data-to-Text – Human Evaluation

Inaccurate

GOLD:

Binignit is a dessert that comes from the Philippines.

PREDICTION:

Philippines is a dessert.

- Zero-shot & few-shot performs comparably
- LLaMA (61%) > Vicuna (48%) > LoRA (17%) > GPT-3.5-Turbo (11%)

© sebis

Data-to-Text – Human Evaluation

Mistranslated

GOLD:

The Mexican based city of Ciudad Ayala in Morelos stands at 1147 metres above sea level and is in the PDT time zone.

PREDICTION:

Ciudad Ayala es parte de Morelos, que es un estado en México. Ciudad Ayala tiene una elevación de 1147 metros sobre el nivel del mar [...]

- Only occurred once for Vicuna zero-shot

© sebis

Data-to-Text – Human Evaluation

Redundant

GOLD:

The Pontiac Rageous, a car with a coupe body style, assembled in Detroit Michigan, was first and last produced in 1997.

PREDICTION:

1997 Pontiac Rageous was produced in Detroit, Michi- gan, in 1997. It was a coupe. It was assembled in 1997. It was produced in 1997 [...] It was a coupe. It was assembled in 1997 [...]

- Mostly an issue of LLaMA zero-shot (23%*)

** Excluding Off-prompt predictions*

© sebis

RQ-3

Data-to-Text – Human Evaluation

Unlexicalized

GOLD:

The Fellowship of the Ring was followed by The Two Towers.

PRED:

The_Fellowship_of_the_Ring was followed by The_Two_Towers.

- Does not occur for LoRA
- Is solved by few-shot prompting for all other models

Appendix: Detailed Results

© sebis

Conclusion

Data-to-Text

- LLMs provide suitable output without task-specific fine-tuning

Semantic Parsing

- More complex compared to data-to-text

→ Task-specific fine-tuning leads to significant improvements

Few-shot prompting

- Reduces errors like: off-prompt, syntax error, formatting issues

Training techniques

- Lead to different error types (i.e. Vicuna: mistranslated, LLaMA: instruction following)

Fine-tuning

- Effective for task-specific generations
- Does not require few-shot examples

Limitations & Future Work

- Only address English texts
  ⇨ Evaluation performance on other languages
- Only considered components in isolation
  ⇨ Develop end-to-end system or plugins
- Model only optimized for a specific task
  ⇨ Fine-tune model on multiple tasks
- Extend evaluation
  ⇨ More LLMs, prompting techniques, datasets, annotators...
  ⇨ Enterprise environment

© sebis

Questions?

References

[1] Thorbecke, C. Google shares lose $100 billion after company's AI chatbot makes an error during demo. Accessed: 2023-10-22.

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. LLaMA: Open and Efficient Foundation Language Models. 2023. arXiv: 2302.13971 [cs.CL].

[3] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. Mar. 2023.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. 2021. arXiv: 2106.09685 [cs.CL].

[5] OpenAI. Model-Overview. Accessed: 2023-10-22.

[6] A. Saha, V. Pahuja, M. Khapra, K. Sankaranarayanan, and S. Chandar. "Complex Sequential Question Answering: Towards Learning to Converse Over Linked Question Answer Pairs with a Knowledge Graph".

[7] L. Perez-Beltrachini, P. Jain, E. Monti, and M. Lapata. *Semantic Parsing for Conversational Question Answering over Knowledge Graphs*. 2023. arXiv: 2301.12217 [cs.CL].

[8] T. Castro Ferreira, C. Gardent, N. Ilinykh, C. van der Lee, S. Mille, D. Moussallem, and A. Shimorina. "The 2020 Bilingual, Bi-Directional WebNLG+ Shared Task: Overview and Evaluation Results (WebNLG+ 2020)"

© sebis

Prof. Dr.

Florian Matthes

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business
Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.tum.de
wwwmatthes.in.tum.de

Method

Semantic Parsing

Sources:

- ACM Digital Library
- IEEE Xplore
- Scopus

Search String:

("semantic parsing" OR "query generation" OR "query creation") AND ("large language model" OR "pretrained language model" OR "pre-trained language model")

Inclusion & Exclusion Criteria:

- Written in English
- Use English texts or datasets
- Published in or after 2020
- Peer-reviewed papers

Data-to-Text Generation

Sources:

- ACM Digital Library
- IEEE Xplore
- Scopus

Search String:

("data-to-text" OR "triples-to-text") AND ("large language model" OR "pretrained language model" OR "pre-trained language model")

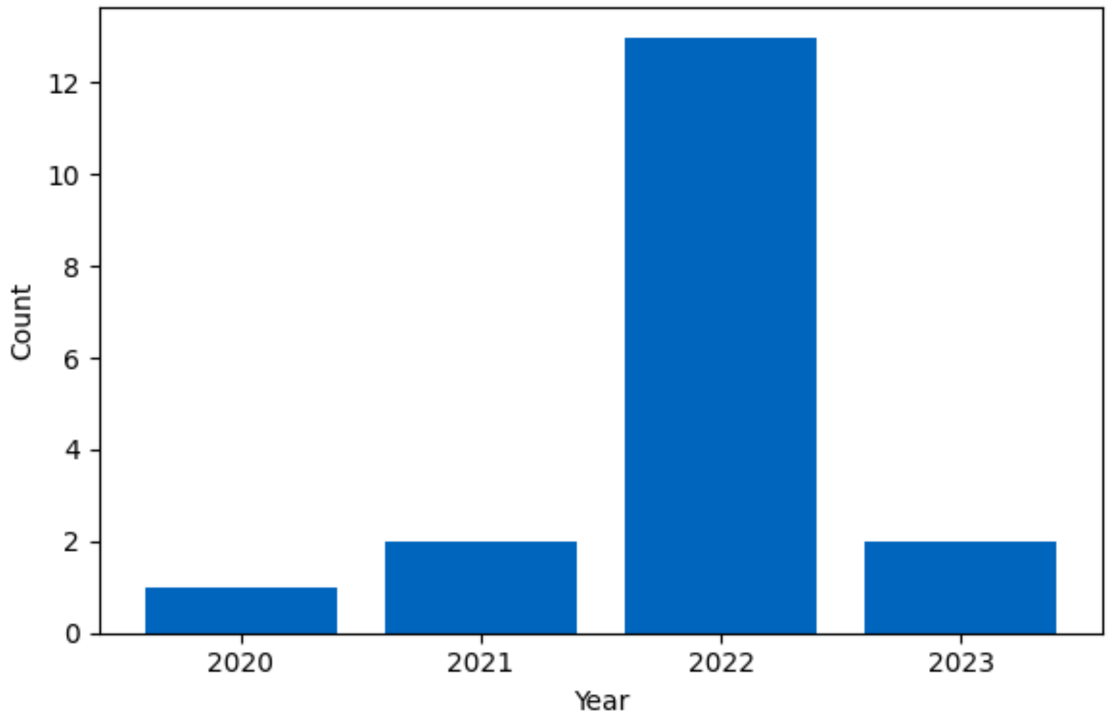
Inclusion & Exclusion Criteria:

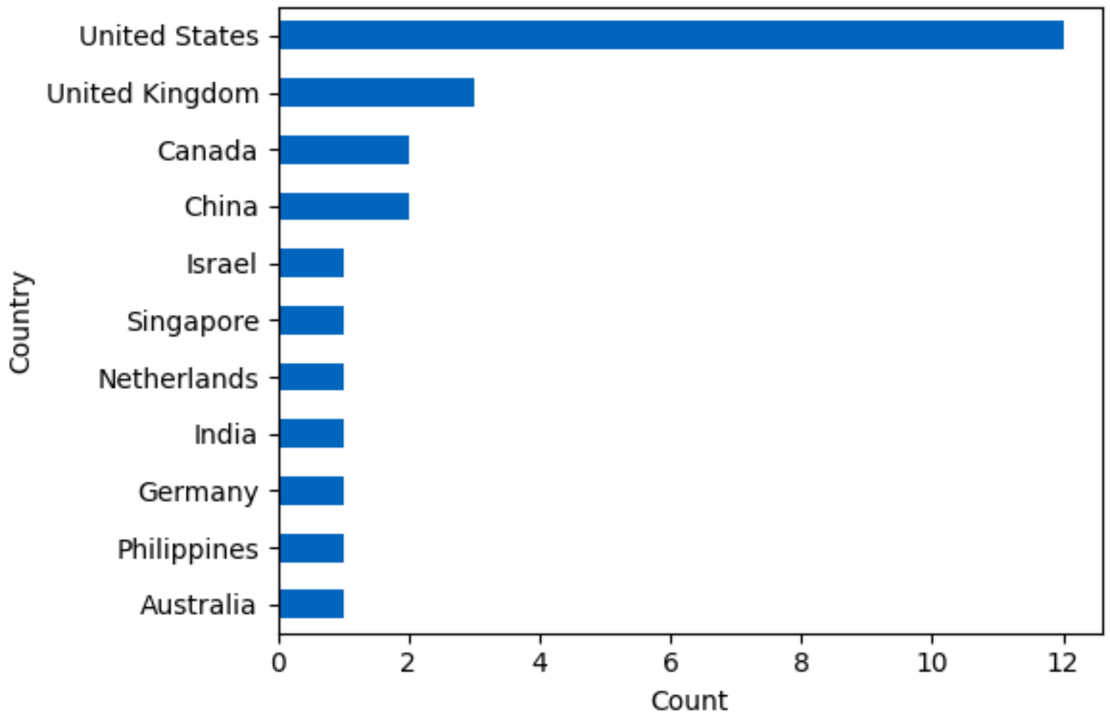
- Written in English
- se English texts or datasets
- Published in or after 2020
- Peer-reviewed papers

© sebis

Findings

Semantic Parsing

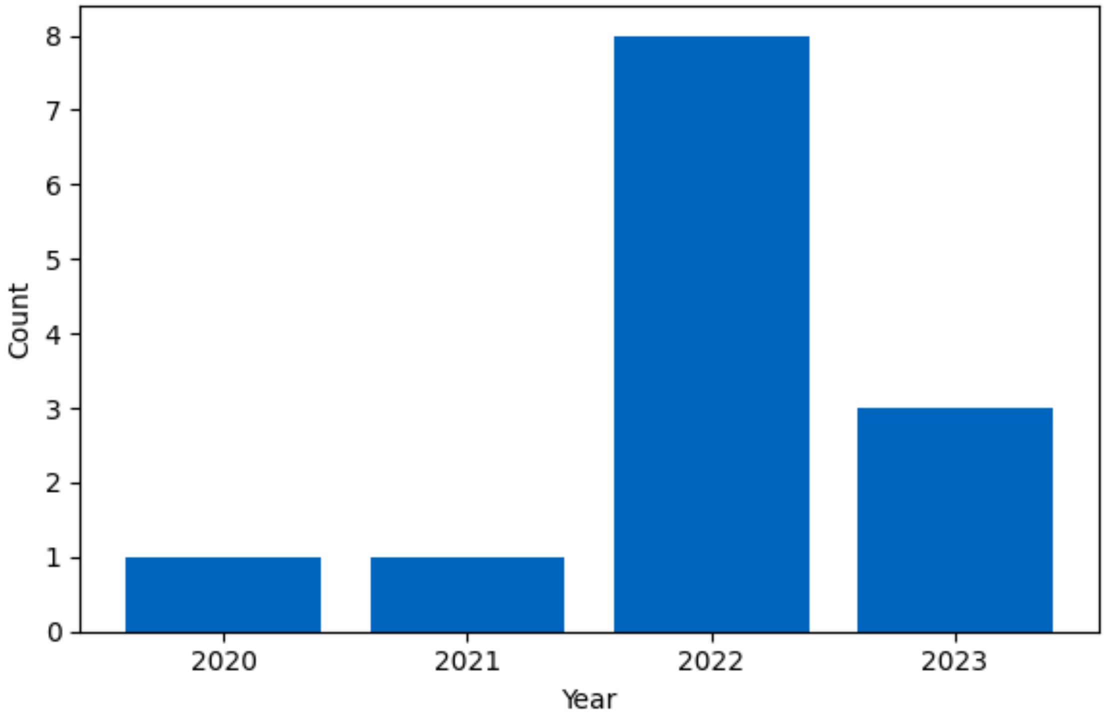


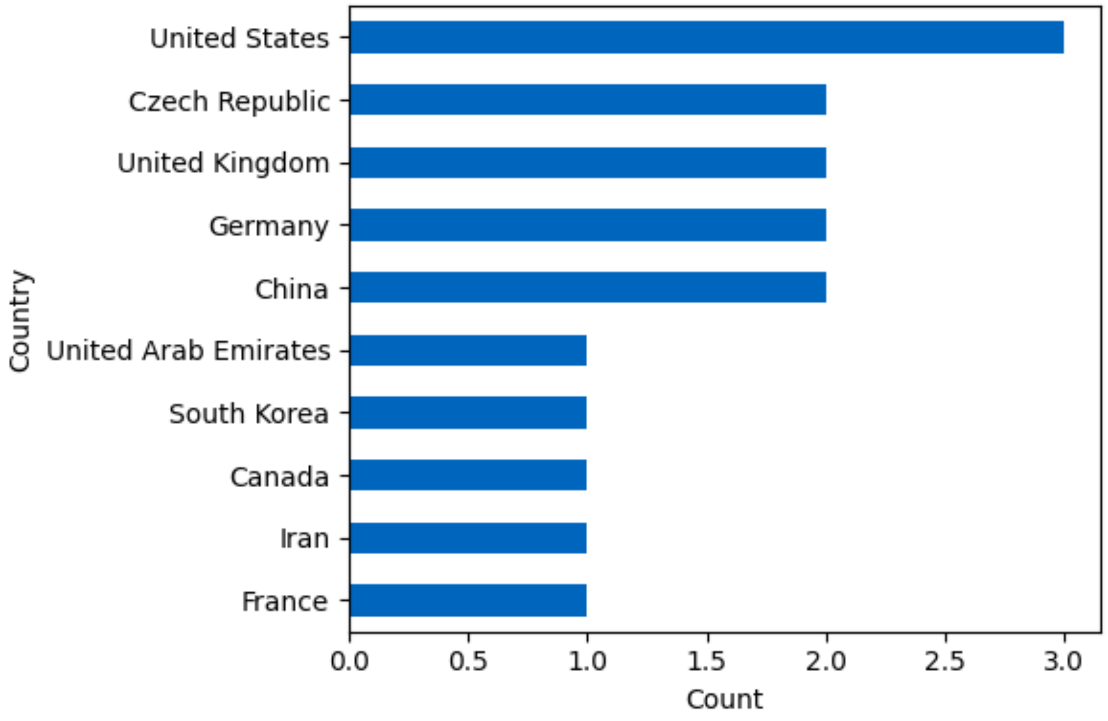


© sebis

Findings

Data-to-Text

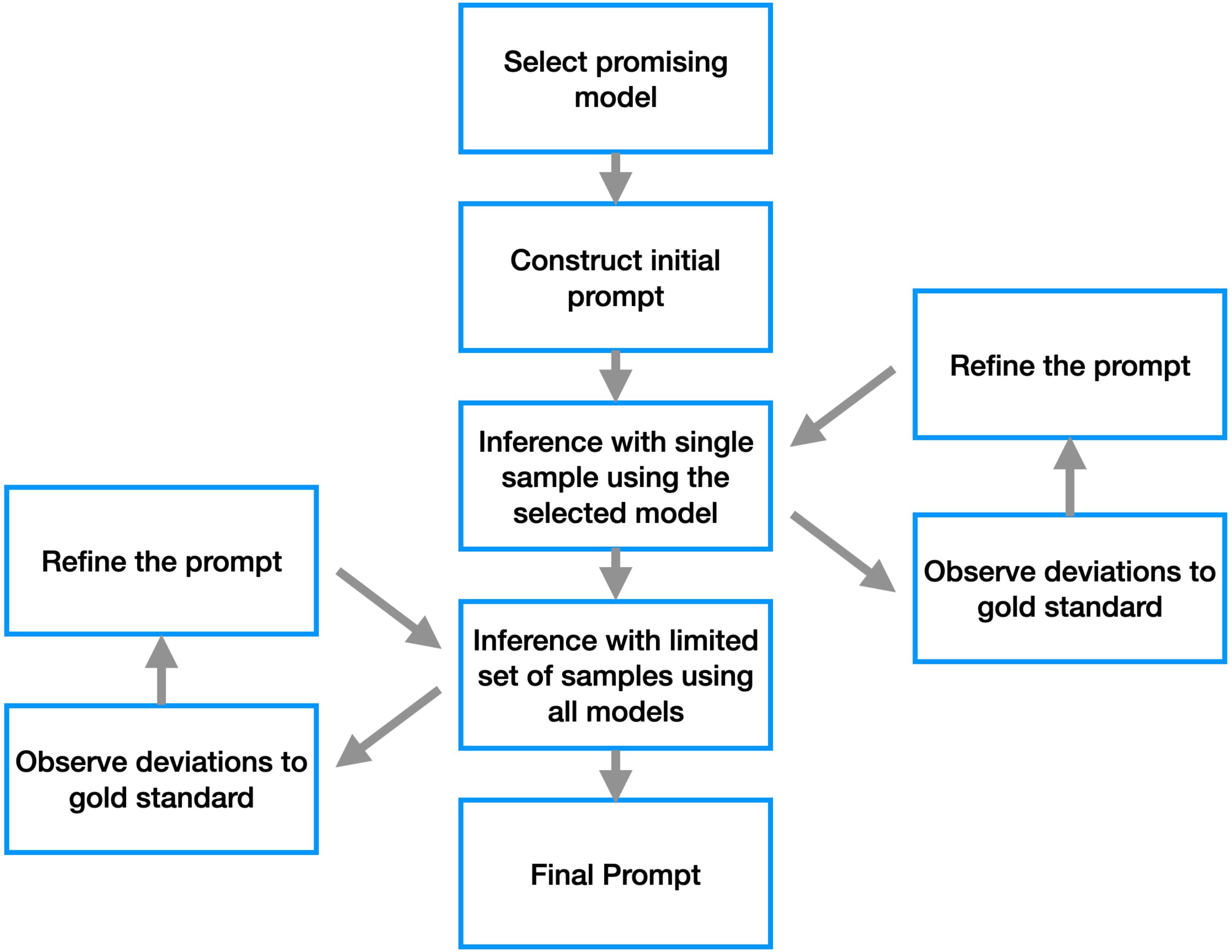




© sebis

RQ-2

Prompt Creation - Method



© sebis

Prompt Creation – Semantic Parsing

Few-shot Template

...

Input question: How many works of art express Michael Jordan or pain ?
Entities: {'Q41421': 'Michael Jordan', 'Q81938': 'pain'}
Relations: {'P180': 'depicts'}
Types: {'Q838948': 'work of art'}

SPARQL query: SELECT (COUNT(DISTINCT ?x) AS ?count) WHERE { { ?x wdt:P180 wd:Q41421 . ?x wdt:P31 wd:Q838948 . } UNION { ?x wdt:P180 wd:Q81938 . ?x wdt:P31 wd:Q838948 . } }

...

© sebis

Prompt Creation – Semantic Parsing

Few-shot Template

...

Conversation history:
USER: Which administrative territory is the native country of Cirilo Villaverde ?
SYSTEM: {’Q241’: ’Cuba’}

Input question: Which is the national anthem of that administrative territory ?
Entities: {’Q241’: ’Cuba’}
Relations: {’P85’: ’anthem’}
Types: {’Q484692’: ’hymn’}

SPARQL query: SELECT ?x WHERE { wd:Q241 wdt:P85 ?x . ?x wdt:P31 wd:Q484692 . }
...

RQ-2

Prompt Creation – Data-to-Text

Few-shot Template

...

Input triples: [{’object’: ’Mike_Mularkey’,’property’: ’coach’,’subject’: ’Ten- nessee_Titans’}]

Output text: Mike Mularkey is the coach of the Tennessee Titans.

...

Few-shot Template

...

Input triples: [{’object’: ’College_of_William_&_Mary’, ’property’: ’owner’, ’sub- ject’: ’Alan_B._Miller_Hall’}, {’object’: ’2009-06-01’, ’property’: ’completionDate’, ’subject’: ’Alan_B._Miller_Hall’}, {’object’: ’101 Ukrop Way’, ’property’: ’address’, ’subject’: ’Alan_B._Miller_Hall’}, {’object’: ’Williamsburg,_Virginia’, ’property’: ’location’, ’subject’: ’Alan_B._Miller_Hall’}, {’object’: ’Robert_A._M._Stern’, ’prop- erty’: ’architect’, ’subject’: ’Alan_B._Miller_Hall’}]

Output text: The Alan B Miller Hall’s location is 101 Ukrop Way, Williams- burg, Virginia. It was designed by Robert A.M. Stern and was completed on 1 June 2009. Its owner is the College of William and Mary.

...

© sebis

Semantic Parsing - Dataset

SPICE


```
{
        "question-type": "Simple Question (Direct)",
        "description": "Simple Question",
        "entities_in_utterance": ["Q131993"],
        "relations": ["P19"],
        "type_list": ["Q852446"],
        "speaker": "USER",
        "utterance": "Which US administrative territory was Wilford Bacon Hoggatt born in ?"
},
{
        "all_entities": ["Q1415"],
        "speaker": "SYSTEM",
        "entities_in_utterance": ["Q1415"],
        "utterance": "Indiana",
        "sparql": "SELECT ?x WHERE { wd:Q131993 wdt:P19 ?x . ?x wdt:P31 wd:Q852446 . }"
},
```

Semantic Parsing – Question Types

Simple Question (Direct):
What is the capital of Germany?

Verification (Boolean) (All):
Is Paris located in Italy?

Quantitative Reasoning (Count) (All)
How many people starred in “Who am I?” or “Sherlock Holms”?

Simple Question (Coreference):
Which genre is associated with it?

Simple Question (Ellipsis):
And which river?

[9]

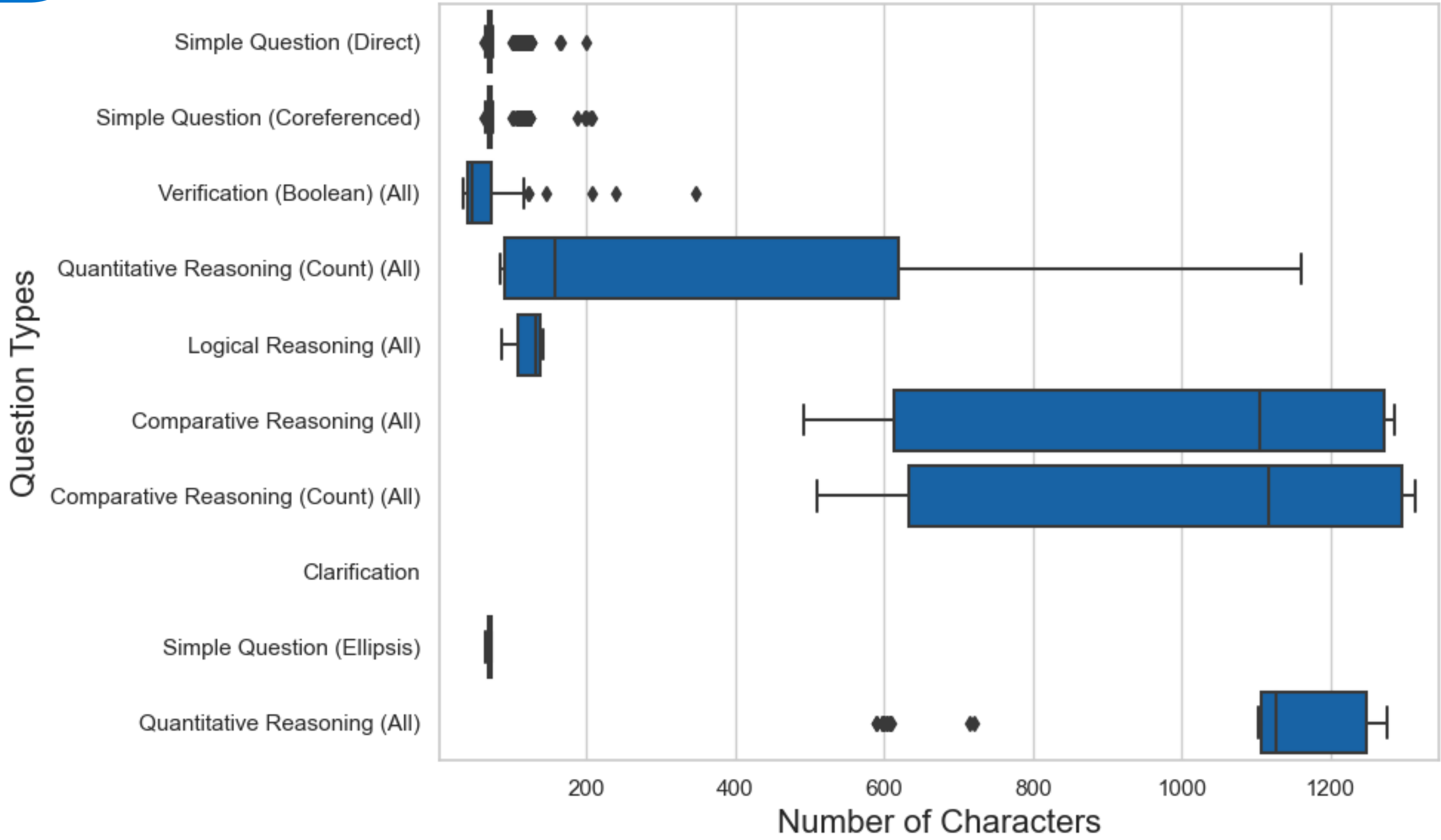
© sebis

Semantic Parsing - Results

© sebis

RQ-3

Semantic Parsing – Query Length



© sebis

RQ-3

Semantic Parsing – Human Evaluation

Issue Type	LLaMA	Vicuna	GPT-3.5-Turbo	LoRA
	relative frequency: zero-shot / few-shot			
Off-prompt	1.00 / 0.10	0.13 / 0.10	0.10 / 0.10	0.10 / 0.10
Cut-off	- / -	- / -	- / -	0.33 * / 0.24 *
Syntax Error	- / 0.16 **	0.71 ** / 0.26 **	0.20 ** / 0.17 **	0.01 ** / 0.10 **
Different Query	- / -	- / -	0.08 / 0.06	0.01 / 0.01
Incorrect Result	- / 0.82 ***	0.97 *** / 0.86 ***	0.69 *** / 0.63 ***	0.12 *** / 0.20 ***
Deviating Entities	- / 0.05 *	0.04 * / 0.02 *	0.06 * / 0.03 *	0.02 * / 0.04 *
Namespace Definition	- / -	0.11 * / -	- / -	- / -
Language Filter	- / -	0.33 * / 0.06 *	0.07 * / 0.02	- / -

* Excluding off-prompt predictions
** Excluding off-prompt and cut-off predictions
*** Excluding off-prompt and syntax error predictions

© sebis

RQ-3

Semantic Parsing – Human Evaluation

Different Query	Deviating SPARQL query returning correct result	GOLD: SELECT ?x WHERE { wd:Q6177791 wdt:P451 ?x . ?x wdt:P31 wd:Q502895 . } PRED: SELECT ?x WHERE { ?x wdt:P451 wd:Q6177791 . ?x wdt:P31 wd:Q502895 . }
Incorrect Result	Valid SPARQL query delivering undesired result	GOLD: SELECT ?x WHERE { wd:Q6177791 wdt:P451 ?x . ?x wdt:P31 wd:Q502895 . } PRED: SELECT ?x WHERE { ?x wdt:P451 ?p . ?p wdt:Q502895 ?type . ?type wdt:commonName ?x . }
Deviating Entities	Prediction uses entity reference not specified in the prompt	GOLD: SELECT DISTINCT ?x WHERE { ?x wdt:P101 ?y . VALUES ?y { wd:Q1622272 wd:Q170790 }. ?x wdt:P31 wd:Q502895 . } } PRED: SELECT ?x WHERE { ?x wdt:P101 wd:Q1622272 . ?x wdt:P101 wd:Q170790 . ?x wdt:P31 wd:Q5 . }
Namespace Definition	Prediction defines namespaces instead of using wd and wdt	GOLD: SELECT ?x WHERE { [...] PRED: PREFIX wdt: <http://www . wikidata . org/entity/> PREFIX wd: <http://www . wikidata . org/prop/direct/> SELECT ?x WHERE { [...]
Language Filter	Predicted query filters the language	GOLD: SELECT ?x WHERE { wd:Q123179 wdt:P69 ?x . ?x wdt:P31 wd:Q163740 . } PRED: SELECT ?x WHERE { ?x wdt:P69 ?y . FILTER (LANG(?y)='en') . } LIMIT 1

Data-to-Text - Dataset

WebNLG


```
"1": {
  "category": "SportsTeam",
  "lexicalisations": [
    {
      "comment": "",
      "lex": "Estádio Municipal Coaracy da Mata Fonseca is the name of the ground of
Agremiação Sportiva Arapiraquense in Arapiraca. Agremiação Sportiva Arapiraquense, nicknamed
\"Alvinegro\", lay in the Campeonato Brasileiro Série C league from Brazil.",
    },
    ...
  ],
  "modifiedtripleset": [
    {
      "object": "Arapiraca",
      "property": "location",
      "subject": "Estádio_Municipal_Coaracy_da_Mata_Fonseca"
    },
    ...
  ],
  "size": "5",
},
```

RQ-3

Data-to-Text - Findings

Model-Prompt	Bleu	Meteor	TER	BERT-Score P	BERT-Score R	BERT-Score F1
Copy-Baseline	0.21	0.02	0.95	0.78	0.81	0.79
LLaMA-7B zero-shot	6.42	0.21	1.03	0.8	0.88	0.84
LLaMA-7B zero-shot-pp	14.21	0.25	0.76	0.88	0.9	0.89
LLaMA-7B few-shot	11.65	0.26	1.03	0.8	0.91	0.85
LLaMA-7B few-shot-pp	37.9	0.36	0.53	0.94	0.94	0.94
Vicuna-7B zero-shot	26.66	0.35	0.68	0.92	0.93	0.92
Vicuna-7B zero-shot-pp	26.66	0.35	0.68	0.92	0.93	0.92
Vicuna-7B few-shot	39.09	0.38	0.64	0.92	0.94	0.93
Vicuna-7B few-shot-pp	43.9	0.39	0.51	0.95	0.95	0.95
GPT-3.5-Turbo zero-shot	41.71	**0.41**	0.56	0.95	0.95	0.95
GPT-3.5-Turbo zero-shot-pp	41.71	**0.41**	0.56	0.95	0.95	0.95
GPT-3.5-Turbo few-shot	39.78	0.4	0.65	0.93	0.95	0.94
GPT-3.5-Turbo few-shot-pp	44.23	**0.41**	0.5	0.95	**0.96**	0.95
LoRA-7B zero-shot	47.25	0.4	0.55	0.93	0.94	0.94
LoRA-7B zero-shot-pp	52.55	**0.41**	**0.42**	**0.96**	**0.96**	**0.96**
LoRA-7B few-shot	47.3	0.4	0.55	0.93	0.94	0.94
LoRA-7B few-shot-pp	**52.89**	**0.41**	**0.42**	**0.96**	**0.96**	**0.96**

RQ-3

Data-to-Text - Human Evaluation

Issue Type	LLaMA	Vicuna	GPT-3.5-Turbo	LoRA
	relative frequency: zero-shot / few-shot			
Off-prompt	0.65 / -	0.27 / -	- / -	- / -
Inaccurate	0.60 * / 0.61	0.41 * / 0.48	0.13 / 0.11	0.19 / 0.17
Mistranslated	- / -	0.01 * / -	- / -	- / -
Redundant	0.23 * / 0.07	0.02 * / -	0.01 / 0.01	- / 0.01
Unlexicalized	0.69 * / -	0.27 * / -	0.07 / -	- / -

* Excluding off-prompt predictions

Prompting Techniques

Zero-Shot

MA Classify the following text into "positive" or "negative" and only return the category.
Text: "Coding is fun."

Positive

Chain-of-Thought (CoT)

MA Lucy has 5 apples. She receives two more from Bob, eats one of them and gives half of the apples that are left to Anna. How many apples are left for Lucy?
Let's think step by step,

Step 1: Lucy has 5 apples.

Step 2: She receives two more from Bob. So now Lucy has 5 + 2 = 7 apples.

Few-Shot

MA Classify the following text into "positive" or "negative" and only return the category.
Text: "Coding is fun."
Category: positive

Text: "Housing prices in Munich are too expensive"
Category: negative

Text: "TUM is an excellent university"
Category:

positive

Other

- System instruction
- Repeating key instructions
- Least-to-Most Prompting
- Mock-Acknowledgement
- Naming the Agent
- ...

© sebis

Use-Cases

Natural Language Generation

Existing Research

- Fusion in Decoder
  - Approach for utilizing the information of multiple input sources to generate a single response
- Retrieval Augmented Generation
  - Combination of information retrieval and response generation

What we want to do

- Retrieval Augmented Generation based on structured data
  - Finetune and test different models
  - Evaluate the different approaches

=> Generates more natural sounding and dynamic responses based on KB data
=> No need for template responses