Polifonia

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

STLAB
SEMANTIC TECHNOLOGY LABORATORY

LREC-COLING 2024

Latent vs Explicit Knowledge Representation: How ChatGPT Answers Questions about Low-Frequency Entities

A. Graciotti[1], V. Presutti[1], R. Tripodi[2]

[1] University of Bologna, 40126 Bologna, Italy
[2] Ca' Foscari University of Venice, Sestiere Dorsoduro, 3246, 30123 Venezia VE

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement GA101004746.
The communication reflects only the author's view and the Research Executive Agency is not responsible for any use that may be made of the information it contains.

Introduction

Problem Statement

Question Answering → LLMs struggle when asked questions about **less popular** factual knowledge [1], as **historical** knowledge.

Benchmark crisis → need for dynamic benchmarks due to LLMs data contamination concerns. [2]

[1] When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories (Mallen et al., ACL 2023)

[2] NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark (Sainz et al., EMNLP 2023)

Introduction

Problem Statement

Question Answering → LLMs struggle when asked questions about **less popular** factual knowledge [1], as **historical** knowledge.

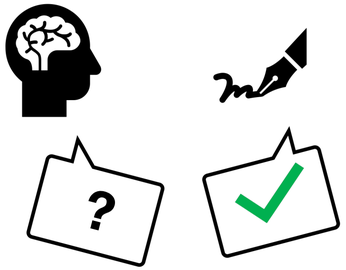
Benchmark crisis → need for dynamic benchmarks due to LLMs **data contamination** concerns. [2]

[1] When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories (Mallen et al., ACL 2023)

[2] NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark (Sainz et al., EMNLP 2023)

Introduction

Contributions



DynaKnowledge [1]: a new dynamic benchmark for free-form QA models

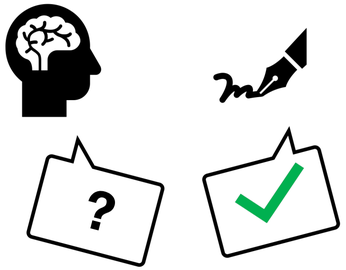
Text2AMR2FRED [2]: a Knowledge Extraction pipeline based on explicit knowledge that can be used for QA

a systematic **comparison** between ChatGPT and the explicit knowledge model centred on entity popularity

[1] https://github.com/polifonia-project/llms-vs-specialised-knowledge

[2] Text2AMR2FRED, a Tool for Transforming Text into RDF/OWL KnowledgeGraphsvia Abstract Meaning Representation (Gangemi et al., ISWC 2023)

Contributions



DynaKnowledge [1]: a new dynamic benchmark for free-form QA models

Text2AMR2FRED [2]: a Knowledge Extraction pipeline based on explicit knowledge that can be used for QA

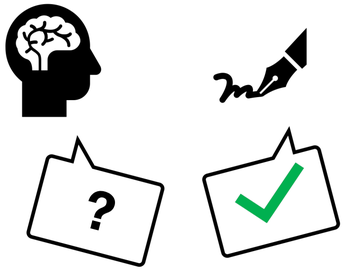
a systematic **comparison** between ChatGPT and the explicit knowledge model centred on entity popularity

[1] https://github.com/polifonia-project/llms-vs-specialised-knowledge

[2] Text2AMR2FRED, a Tool for Transforming Text into RDF/OWL KnowledgeGraphsvia Abstract Meaning Representation (Gangemi et al., ISWC 2023)

Introduction

Contributions



DynaKnowledge [1]: a new dynamic benchmark for free-form QA models

---

Text2AMR2FRED [2]: a Knowledge Extraction pipeline based on explicit knowledge that can be used for QA

a systematic **comparison** between ChatGPT and the explicit knowledge model centred on entity popularity

[1] https://github.com/polifonia-project/llms-vs-specialised-knowledge

[2] Text2AMR2FRED, a Tool for Transforming Text into RDF/OWL KnowledgeGraphsvia Abstract Meaning Representation (Gangemi et al., ISWC 2023)

Methodology

Objective → comparing the performance of explicit and latent knowledge models on the QA task

Steps → collecting a new benchmark for the task;
selecting two representative models for latent and explicit knowledge representation

Objective → comparing the performance of explicit and latent knowledge models on the QA task

Steps → collecting a new benchmark for the task;
selecting two representative models for latent and explicit knowledge representation

Methodology

DynaKnowledge: a new Dynamic Benchmark for QA

Composition → 82 *question*, *answer*, *provenance* samples centred on historical characters' biographies (**50-50 gender* balance**)

ID	Gender	Analyst's Question	Analyst's Answer	Sentence from Wikipedia containing the answer
18	F	How long was Teresina Brambilla's career as a musician?	25 years	Teresa "Teresina" Brambilla (15 April 1845 – 1 July 1921) was an Italian soprano who sang in the major opera houses of Europe in a career spanning 25 years.
41	F	Which female colleague did Giuseppina Ronzi de Begnis argue with during the rehearsals of Maria Stuarda?	Anna Del Sere	Ronzi was also known for her capricious attitudes and for having confrontations and arguments with female colleagues, including the famous altercation with Anna Del Sere during the rehearsals of Maria Stuarda.
5	M	How old was Wolfgang Amadeus Mozart when he started to compose?	5 years old	Already competent on keyboard and violin, he composed from the age of five and performed before European royalty.

*In our study, we restrict to binary gender categories, which, although not reflecting real-world diversity, let us move the first steps towards the definition of our method.

DynaKnowledge: a new Dynamic Benchmark for QA

Composition → 82 *question*, *answer*, *provenance* samples centred on historical characters' biographies (**50-50 gender* balance**)

ID	Gender	Analyst's Question	Analyst's Answer	Sentence from Wikipedia containing the answer
18	F	How long was Teresina Brambilla's career as a musician?	25 years	Teresa "Teresina" Brambilla (15 April 1845 – 1 July 1921) was an Italian soprano who sang in the major opera houses of Europe in a career spanning 25 years.
41	F	Which female colleague did Giuseppina Ronzi de Begnis argue with during the rehearsals of Maria Stuarda?	Anna Del Sere	Ronzi was also known for her capricious attitudes and for having confrontations and arguments with female colleagues, including the famous altercation with Anna Del Sere during the rehearsals of Maria Stuarda.
5	M	How old was Wolfgang Amadeus Mozart when he started to compose?	5 years old	Already competent on keyboard and violin, he composed from the age of five and performed before European royalty.

*In our study, we restrict to binary gender categories, which, although not reflecting real-world diversity, let us move the first steps towards the definition of our method.

DynaKnowledge: a new Dynamic Benchmark for QA

Composition → 82 *question*, *answer*, *provenance* samples centred on historical characters' biographies (**50-50 gender* balance**)

ID	Gender	Analyst's Question	Analyst's Answer	Sentence from Wikipedia containing the answer
18	F	How long was Teresina Brambilla's career as a musician?	25 years	Teresa "Teresina" Brambilla (15 April 1845 – 1 July 1921) was an Italian soprano who sang in the major opera houses of Europe in a career spanning 25 years.
41	F	Which female colleague did Giuseppina Ronzi de Begnis argue with during the rehearsals of Maria Stuarda?	Anna Del Sere	Ronzi was also known for her capricious attitudes and for having confrontations and arguments with female colleagues, including the famous altercation with Anna Del Sere during the rehearsals of Maria Stuarda.
5	M	How old was Wolfgang Amadeus Mozart when he started to compose?	5 years old	Already competent on keyboard and violin, he composed from the age of five and performed before European royalty.

*In our study, we restrict to binary gender categories, which, although not reflecting real-world diversity, let us move the first steps towards the definition of our method.

Methodology

DynaKnowledge: a new Dynamic Benchmark for QA

Source → a corpus of music historical periodicals (1823 to 1900) [1]

for Rubini, was well sustained by a young singer of the name of Gennero: the Brambilla was also much applauded in this opera.

Recit. ed Aria, "*Di tanti palpiti*," Madame Ronzi De Begnis, *(Tancredi)* - - - - - *Rossini.*
Finale to the 1st Act of *Don Giovanni*, Mrs. Salmon,

Yesterday the *Requiem* of Mozart was performed as a mass for the deceased †, in the church of the Augustine

Teresina Brambilla

Portrait of Brambilla by Ernesto Fontana [Wikidata]

Portrait of Giuseppina de Begnis as Fatima in Rossini's oratorio *Pietro L' Eremita*, London 1828, by Alfred Chalon

Wolfgang Amadeus Mozart

Portrait, c. 1781

[1] The *Periodicals* module of the Polifonia Textual Corpus (https://github.com/polifonia-project/Polifonia-Corpus)

Methodology

Models

Latent Knowledge Representation:
ChatGPT (GPT3.5) [1]

Explicit Knowledge Representation:
Text2AMR2FRED

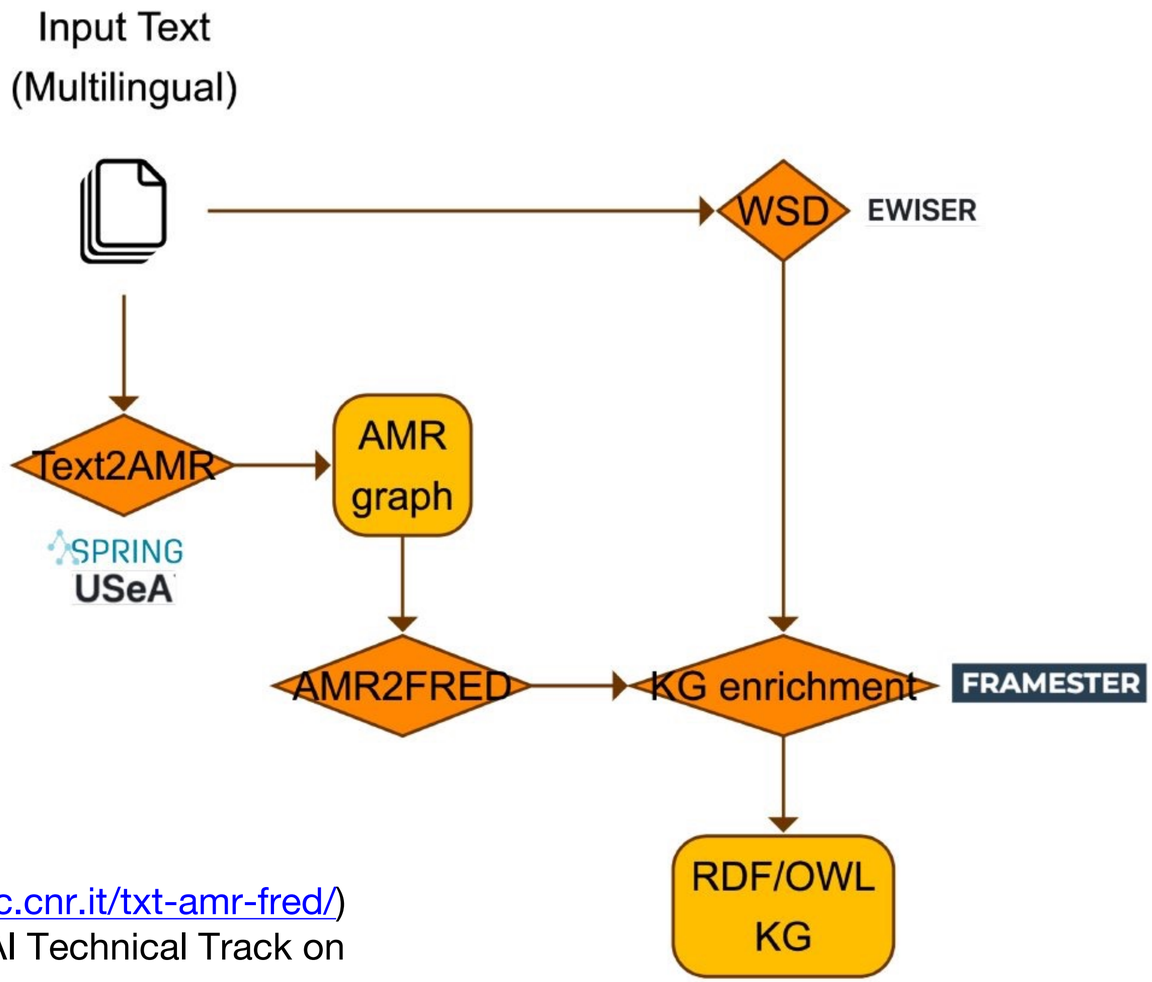
[1] GPT-3.5 (Default) model was accessed via its free online version (https://chat.openai.com/) from August 4th to August 22nd, 2023

Methodology

Models

Latent Knowledge Representation:
ChatGPT (GPT3.5)

Explicit Knowledge Representation:
Text2AMR2FRED [2] [3]



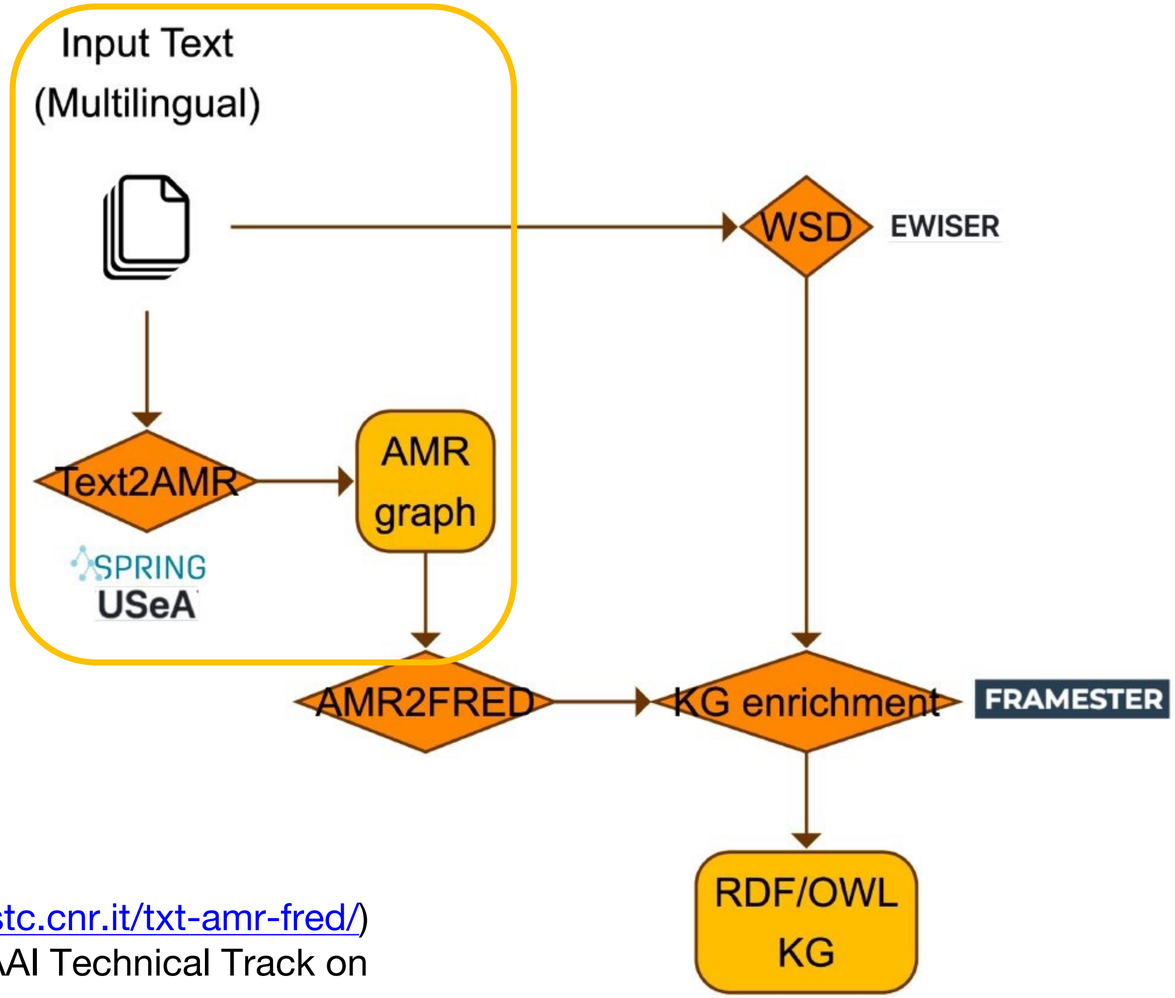
[2] Text2AMR2FRED was accessed via its WebApp (https://arco.istc.cnr.it/txt-amr-fred/)
[3] Text2AMR parsing is enabled by SPRING (Bevilacqua et al., AAAI Technical Track on Speech and Natural Language Processing I 2021)

Methodology

Models

Latent Knowledge Representation:
ChatGPT (GPT3.5)

Explicit Knowledge Representation:
Text2AMR2FRED [2] [3]

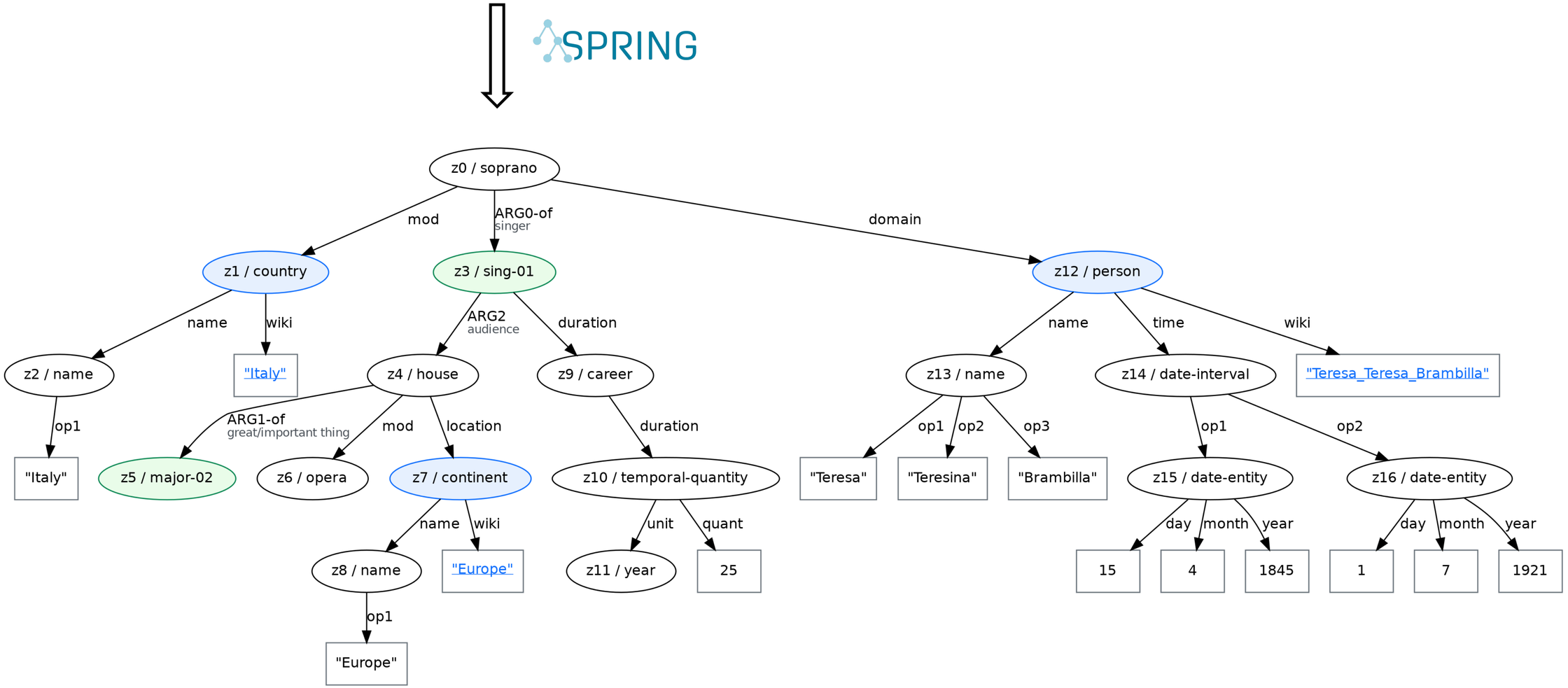


[2] Text2AMR2FRED was accessed via its WebApp (https://arco.istc.cnr.it/txt-amr-fred/)
[3] Text2AMR parsing is enabled by SPRING (Bevilacqua et al., AAAI Technical Track on Speech and Natural Language Processing I 2021)

Methodology

Models

Teresa "Teresina" Brambilla (15 April 1845 – 1 July 1921) was an Italian soprano who sang in the major opera houses of Europe in a career spanning 25 years.



Evaluation

ChatGPT Answer correctness

ChatGPT (GPT3.5)

PROMPT: How long was Teresina Brambilla career as a musician? Give a straight answer .

ChatGPT (3.5) ANSWER: As of my last knowledge update in September 2021, there is no widely known or documented musician by the name of Teresina Brambilla in the public domain. […]

ANALYST'S ASSESSMENT: *Answer not given*

ID	Gender	Analyst's Question	Analyst's Answer	Sentence from Wikipedia containing the answer
18	F	How long was Teresina Brambilla's career as a musician?	25 years	Teresa "Teresina" Brambilla (15 April 1845 – 1 July 1921) was an Italian soprano who sang in the major opera houses of Europe in a career spanning 25 years.

Evaluation

ChatGPT Answer correctness

ChatGPT (GPT3.5)

PROMPT: How long was Teresina Brambilla career as a musician? Give a straight answer .

ChatGPT (3.5) ANSWER: As of my last knowledge update in September 2021, there is no widely known or documented musician by the name of Teresina Brambilla in the public domain. […]

ANALYST'S ASSESSMENT: *Answer not given*

ID	Gender	Analyst's Question	Analyst's Answer	Sentence from Wikipedia containing the answer
18	F	How long was Teresina Brambilla's career as a musician?	25 years	Teresa "Teresina" Brambilla (15 April 1845 – 1 July 1921) was an Italian soprano who sang in the major opera houses of Europe in a career spanning 25 years.

Evaluation

ChatGPT Answer correctness

ChatGPT (GPT3.5)

PROMPT: Which female colleague did Giuseppina Ronzi de Begnis argue with during the rehearsals of Maria Stuarda? Give a straight answer.

ChatGPT (3.5) ANSWER: Giuseppina Ronzi de Begnis argued with Maria Malibran during the rehearsals of Maria Stuarda.

ANALYST'S ASSESSMENT: *Wrong*

ID	Gender	Analyst's Question	Analyst's Answer	Sentence from Wikipedia containing the answer
41	F	Which female colleague did Giuseppina Ronzi de Begnis argue with during the rehearsals of Maria Stuarda?	Anna Del Sere	Ronzi was also known for her capricious attitudes and for having confrontations and arguments with female colleagues, including the famous altercation with Anna Del Sere during the rehearsals of Maria Stuarda.

Evaluation

ChatGPT Answer correctness

ChatGPT (GPT3.5)

PROMPT: Which female colleague did Giuseppina Ronzi de Begnis argue with during the rehearsals of Maria Stuarda? Give a straight answer.

ChatGPT (3.5) ANSWER: Giuseppina Ronzi de Begnis argued with Maria Malibran during the rehearsals of Maria Stuarda.

ANALYST'S ASSESSMENT: *Wrong*

ID	Gender	Analyst's Question	Analyst's Answer	Sentence from Wikipedia containing the answer
41	F	Which female colleague did Giuseppina Ronzi de Begnis argue with during the rehearsals of Maria Stuarda?	Anna Del Sere	Ronzi was also known for her capricious attitudes and for having confrontations and arguments with female colleagues, including the famous altercation with Anna Del Sere during the rehearsals of Maria Stuarda.

Evaluation

ChatGPT Answer correctness

ChatGPT (GPT3.5)

PROMPT: How old was Wolfgang Amadeus Mozart when he started to compose?

ChatGPT (3.5) ANSWER: Wolfgang Amadeus Mozart started composing at the age of 5.

ANALYST'S ASSESSMENT: *Correct*

ID	Gender	Analyst's Question	Analyst's Answer	Sentence from Wikipedia containing the answer
5	M	How old was Wolfgang Amadeus Mozart when he started to compose?	5 years old	Already competent on keyboard and violin, he composed from the age of five and performed before European royalty.

Evaluation

ChatGPT Answer correctness

ChatGPT (GPT3.5)

PROMPT: How old was Wolfgang Amadeus Mozart when he started to compose?

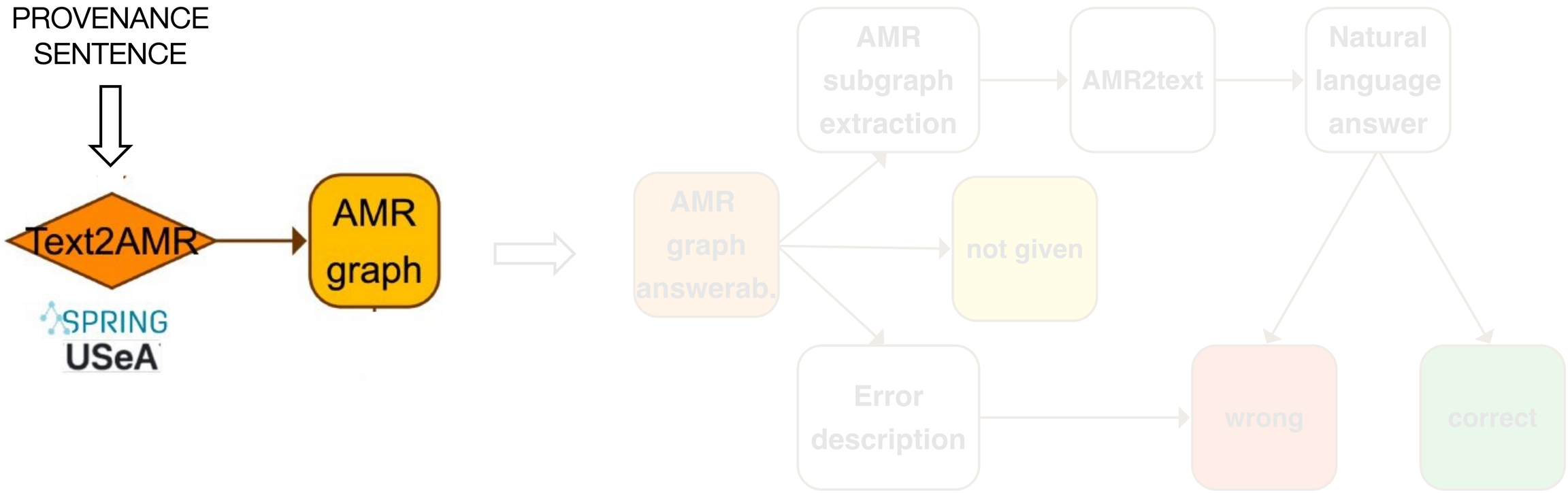
ChatGPT (3.5) ANSWER: Wolfgang Amadeus Mozart started composing at the age of 5.

ANALYST'S ASSESSMENT: *Correct*

ID	Gender	Analyst's Question	Analyst's Answer	Sentence from Wikipedia containing the answer
5	M	How old was Wolfgang Amadeus Mozart when he started to compose?	5 years old	Already competent on keyboard and violin, he composed from the age of five and performed before European royalty.

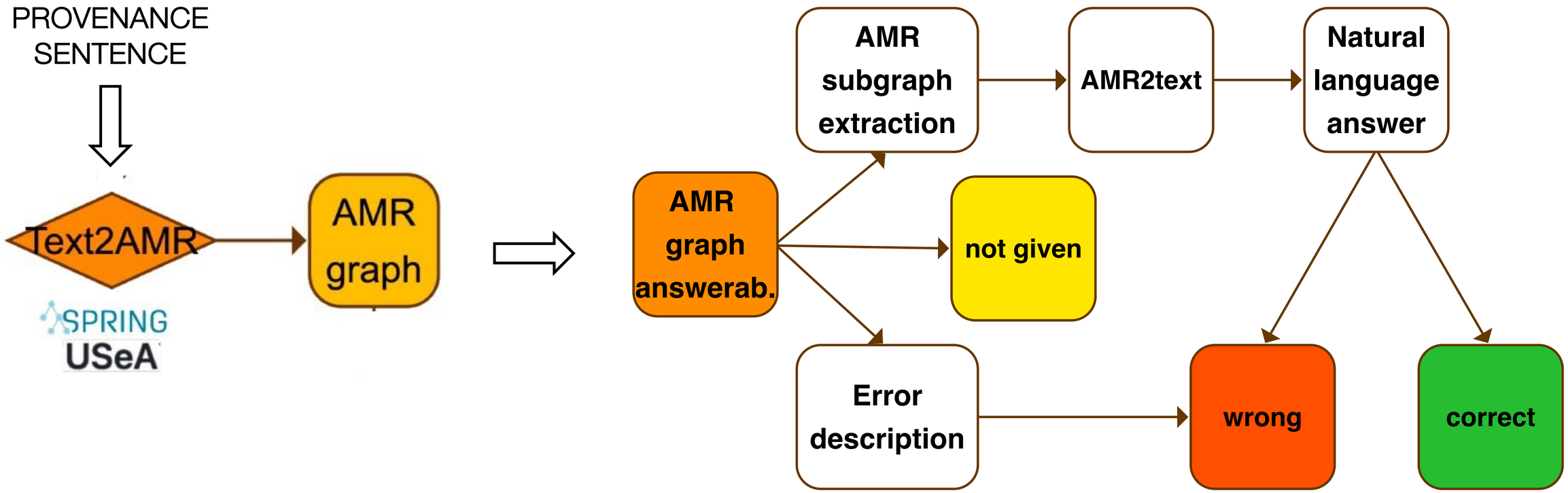
Evaluation

AMR graph answerability



Evaluation

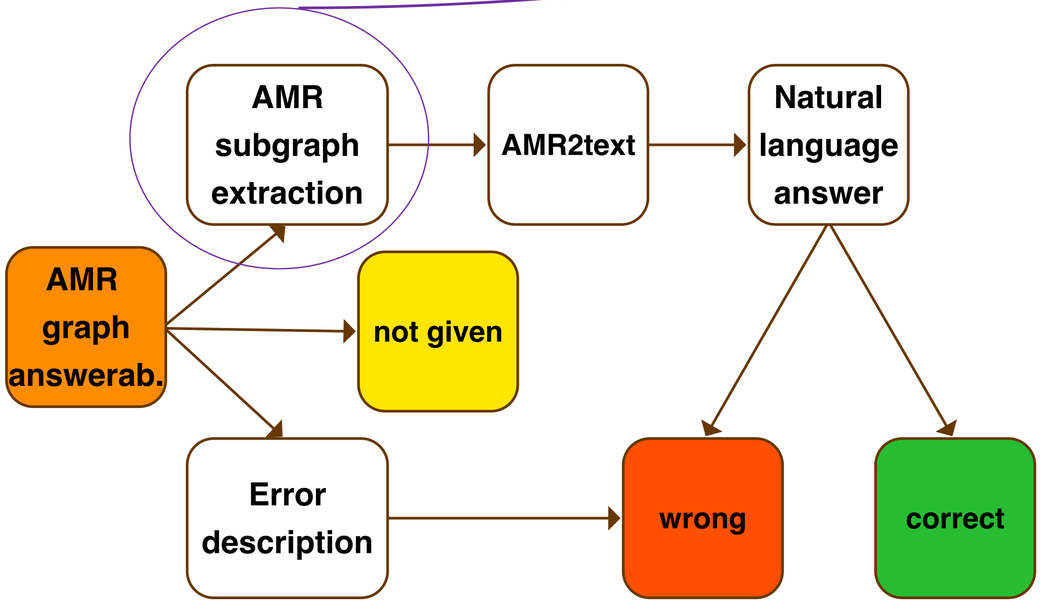
AMR graph answerability

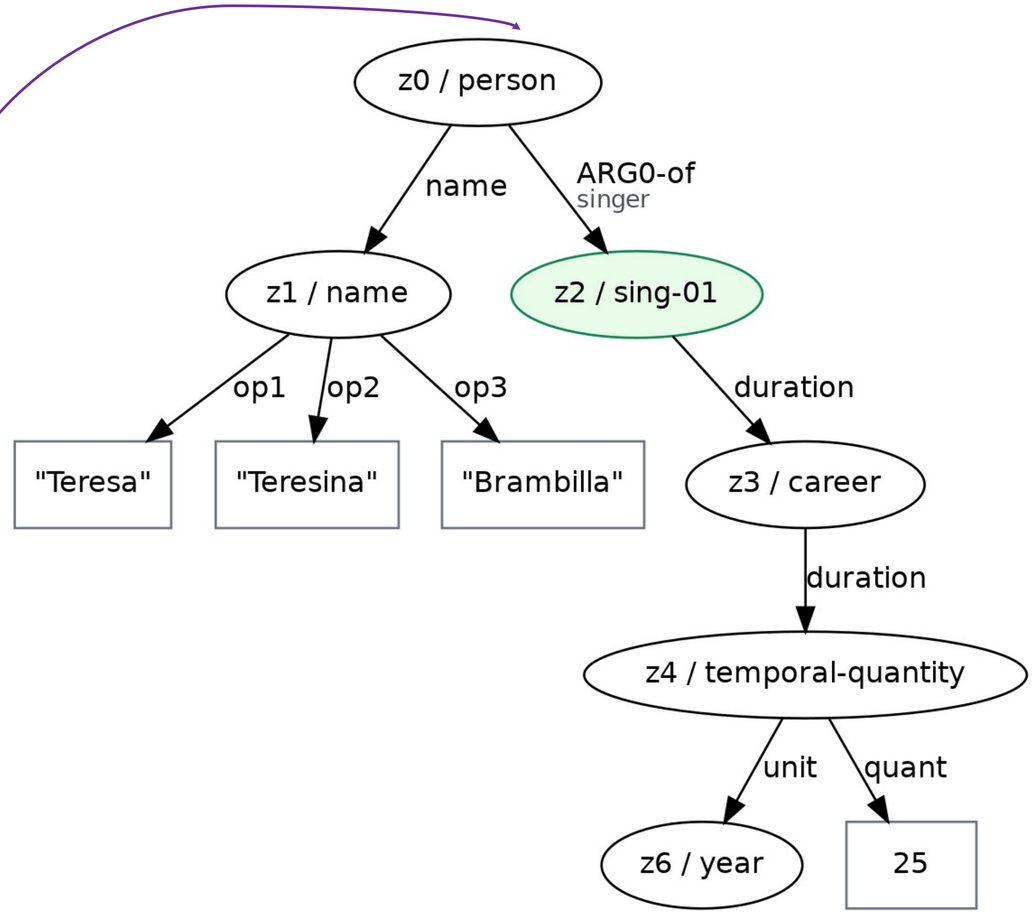


Evaluation

AMR graph answerability

Question: *How long was Teresina Brambilla's career as a musician?*

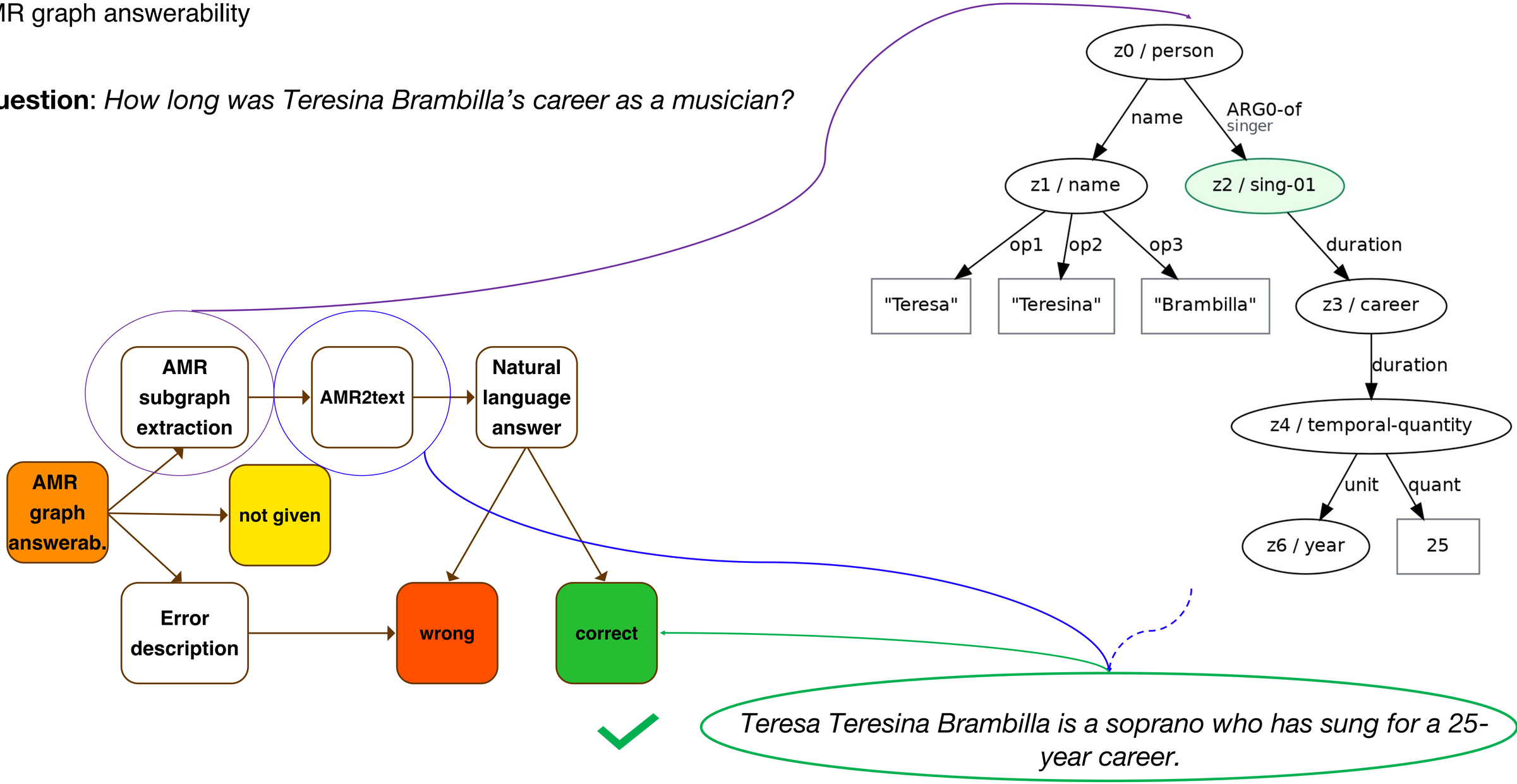




Evaluation

AMR graph answerability

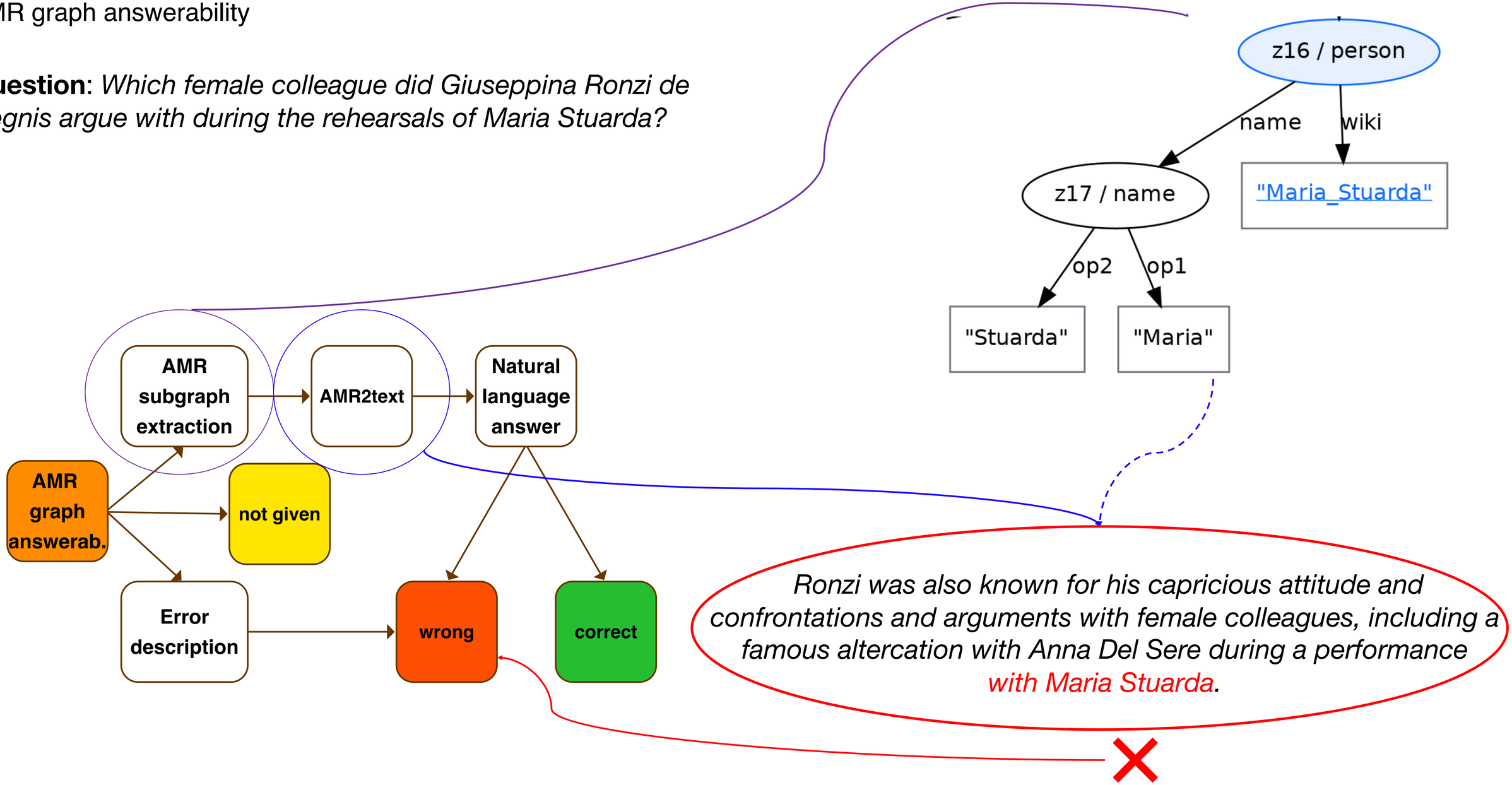
Question: *How long was Teresina Brambilla's career as a musician?*



Evaluation

AMR graph answerability

Question: *Which female colleague did Giuseppina Ronzi de Begnis argue with during the rehearsals of Maria Stuarda?*



Evaluation

Results and Analysis

NE's gender	ChatGPT			AMR		
	P	R	F1	P	R	F1
F	0, 22	0, 31	0, 26	0, 51	0, 51	0, 51
M	0, 54	0, 58	0, 56	0, 49	0, 50	0, 49
Total	0, 38	0, 46	0, 42	**0, 50**	**0, 50**	**0, 50**

$$P = \frac{correct}{|samples|}, \; R = \frac{correct}{|samples| - \text{not given}},$$

and

$$F1 = \frac{2 * P * R}{P + R}.$$

Evaluation

Results and Analysis

NE's gender	ChatGPT			AMR		
	P	R	F1	P	R	F1
F	0, 22	0, 31	0, 26	0, 51	0, 51	0, 51
M	0, 54	0, 58	0, 56	0, 49	0, 50	0, 49
Total	0, 38	0, 46	0, 42	**0, 50**	**0, 50**	**0, 50**

$$P = \frac{correct}{|samples|}, \; R = \frac{correct}{|samples| - \text{not given}},$$

and

$$F1 = \frac{2 * P * R}{P + R}.$$

Evaluation

Results and Analysis

NE's gender	ChatGPT			AMR		
	P	R	F1	P	R	F1
F	0, 22	0, 31	0, 26	0, 51	0, 51	0, 51
M	0, 54	0, 58	0, 56	0, 49	0, 50	0, 49
Total	0, 38	0, 46	0, 42	**0, 50**	**0, 50**	**0, 50**

$$P = \frac{correct}{|samples|}, \; R = \frac{correct}{|samples| - \text{not given}},$$

and

$$F1 = \frac{2 * P * R}{P + R}.$$

Evaluation

Popularity and gender effect

Popularity → each named entity's Wikidata identifier (QID) frequency of occurrence as an internal link in Wikipedia [1].

NE's gender	ChatGPT answer						AMR graph answerability							
	#yes		#no		#not given		#yes		#no		#not given		#Total	
	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev
F	76	126	37	93	4	3	42	96	30	83,5	N/A	N/A	36	89
M	869	1579	201	600	6	3	776	1627	338.5	739	61	N/A	545	1253
All	639	1372	110	407	4	3	400	1183	184	542	61	N/A	291	919

[1] We used the enwiki-20220120 dump.

Evaluation

Popularity and gender effect

Popularity → each named entity's Wikidata identifier (QID) frequency of occurrence as an internal link in Wikipedia [1].

NE's gender	ChatGPT answer						AMR graph answerability							
	#yes		#no		#not given		#yes		#no		#not given		#Total	
	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev
F	76	126	37	93	4	3	42	96	30	83,5	N/A	N/A	36	89
M	869	1579	201	600	6	3	776	1627	338.5	739	61	N/A	545	1253
All	639	1372	110	407	4	3	400	1183	184	542	61	N/A	291	919

[1] We used the enwiki-20220120 dump.

Evaluation

Popularity and gender effect

Popularity → each named entity's Wikidata identifier (QID) frequency of occurrence as an internal link in Wikipedia [1].

NE's gender	ChatGPT answer						AMR graph answerability						#Total	
	#yes		#no		#not given		#yes		#no		#not given			
	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev
F	76	126	37	93	4	3	42	96	30	83,5	N/A	N/A	36	89
M	869	1579	201	600	6	3	776	1627	338.5	739	61	N/A	545	1253
All	639	1372	110	407	4	3	400	1183	184	542	61	N/A	291	919

[1] We used the enwiki-20220120 dump.

Evaluation

Popularity and gender effect

Popularity → each named entity's Wikidata identifier (QID) frequency of occurrence as an internal link in Wikipedia [1].

NE's gender	ChatGPT answer						AMR graph answerability							
	#yes		#no		#not given		#yes		#no		#not given		#Total	
	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev	Avg	StdDev
F	76	126	37	93	4	3	42	96	30	83,5	N/A	N/A	36	89
M	869	1579	201	600	6	3	776	1627	338.5	739	61	N/A	545	1253
All	639	1372	110	407	4	3	400	1183	184	542	61	N/A	291	919

[1] We used the enwiki-20220120 dump.

Evaluation

Popularity and gender effect

NE's feature	Spearman correlation	
	ChatGPT's answer assessment	AMR graph's answerability
Popularity	0.48	−0.07
Gender	0.33	−0.02

Evaluation

Popularity and gender effect

NE's feature	Spearman correlation	
	ChatGPT's answer assessment	AMR graph's answerability
Popularity	0.48	−0.07
Gender	0.33	−0.02

Evaluation

Popularity and gender effect

NE's feature	Spearman correlation	
	ChatGPT's answer assessment	AMR graph's answerability
Popularity	0.48	−0.07
Gender	0.33	−0.02

Evaluation

Popularity and gender effect

NE's feature	Spearman correlation	
	ChatGPT's answer assessment	AMR graph's answerability
Popularity	0.48	−0.07
Gender	0.33	−0.02

Evaluation

Popularity and gender effect

NE's feature	Spearman correlation	
	ChatGPT's answer assessment	AMR graph's answerability
Popularity	0.48	−0.07
Gender	0.33	−0.02

Conclusion

Popularity matters → ChatGPT struggles to answer questions related to less popular entities, which in our benchmark are female.

→ A **simpler knowledge extraction pipeline** demonstrated to be more robust on the variation of named entities' features such as gender and popularity.

Conclusion

Popularity matters → ChatGPT struggles to answer questions related to less popular entities, which in our benchmark are female.

→ A simpler knowledge extraction pipeline demonstrated to be more robust on the variation of named entities' features such as gender and popularity.

Future work

Expand DynaKnowledge dynamically, following current trends in benchmarking [1] NLP models.

Transforming the AMR graphs into OWL-compliant RDF KGs to **automate answerability assessment** through structured interrogations.

Exploiting the KGs to augment LLMs prompts and improve QA performance in a retrieval-augmented generation [2] framework.

[1] Dynabench: Rethinking Benchmarking in NLP (Kiela et al., NAACL 2021)

[2] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (Lewis et al., EMNLP 2023)

Future work

Expand DynaKnowledge dynamically, following current trends in benchmarking [1] NLP models.

Transforming the AMR graphs into OWL-compliant RDF KGs to **automate answerability assessment** through structured interrogations.

Exploiting the KGs to augment LLMs prompts and improve QA performance in a retrieval-augmented generation [2] framework.

[1] Dynabench: Rethinking Benchmarking in NLP (Kiela et al., NAACL 2021)

[2] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (Lewis et al., EMNLP 2023)

Future work

Expand DynaKnowledge dynamically, following current trends in benchmarking [1] NLP models.

Transforming the AMR graphs into OWL-compliant RDF KGs to **automate answerability assessment** through structured interrogations.

Exploiting the KGs to augment LLMs prompts and improve QA performance in a retrieval-augmented generation [2] framework.

[1] Dynabench: Rethinking Benchmarking in NLP (Kiela et al., NAACL 2021)

[2] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (Lewis et al., EMNLP 2023)

Polifonia

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

STLAB
SEMANTIC TECHNOLOGY LABORATORY

LREC-COLING 2024

Thank you for your attention!

Speaker:
Arianna Graciotti

arianna.graciotti@unibo.it

https://www.unibo.it/sitoweb/arianna.graciotti
https://www.linkedin.com/in/arianna-graciotti/

https://github.com/polifonia-project/llms-vs-specialised-knowledge