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

Arianna Graciotti*, Valentina Presutti*, Rocco Tripodi†
*University of Bologna, †Ca' Foscari University of Venice
*Via Zamboni 33, 40126, Bologna, Italy, †Sestiere Dorsoduro 3246, 30123 Venezia, Italy
*{valentina.presutti, arianna.graciotti}@unibo.it
†rocco.tripodi@unive.it

Abstract

In this paper, we present an evaluation of two different approaches to the free-form Question Answering (QA) task. The main difference between the two approaches is that one is based on latent representations of knowledge, and the other uses explicit knowledge representation. For the evaluation, we developed DynaKnowledge, a new benchmark composed of questions concerning Wikipedia low-frequency entities. We wanted to ensure, on the one hand, that the questions are answerable and, on the other, that the models can provide information about very specific facts. The evaluation that we conducted highlights that the proposed benchmark is particularly challenging. The best model answers correctly only on 50% of the questions. Analysing the results, we also found that ChatGPT shows low reliance on low-frequency entity questions, manifesting a popularity bias. On the other hand, a simpler model based on explicit knowledge is less affected by this bias. With this paper, we want to provide a living benchmark for open-form QA to test knowledge and latent representation models on a dynamic benchmark.

Keywords: Question Answering, Knowledge Representation, LLMs

1. Introduction

The introduction of the transformer architecture (Vaswani et al., 2017) revolutionised Natural Language Processing (NLP). It brought a new paradigm in the field with the development of Large Language Models (LLMs), such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020), pre-trained on plain text and specialized on defined tasks. This new paradigm introduced a new way to perform prediction tasks in machine learning, in which instead of modelling directly the probability of an output label y given an input x, it provides for the modification of the input x using a template and prompting the LLM to fill it. The desired output y can be then extracted or inferred from the LLM's response (Liu et al., 2023). A seminal work in this area has been proposed by Raffel et al. (2020), introducing a unified framework that converts text-based language problems into a text-to-text format.

The task of Question Answering (QA) has also been influenced by this new paradigm. Petroni et al. (2019), starting from the assumption that LLMs can learn relational knowledge in the same way they learn linguistic knowledge, proposed an approach to extract this knowledge using *fill-in-the-blank* cloze statements. This line of research paved the way for the development of models that use the latent representations of LLMs as Knowledge Bases (KBs) (Petroni et al., 2019) and use latent representations and KBs in conjunction to approach knowledge-intensive tasks (Lewis et al., 2020; Guu et al., 2020). However, the performance of such models on open-domain QA is still quite low (Siriwardhana et al., 2023).

A related and more general line of research in this context culminated with the development of conversational agents such as ChatGPT (OpenAI, 2022). These models can interact with humans and provide answers and solutions to different tasks prompted by the users. If, from one side, the popularity of this kind of models is increasing constantly, the research community has raised two main concerns to them. The first one is related to the LLMs' tendency to *hallucinate* (Ji et al., 2023) with the consequent generation of biased or harmful content (Bender et al., 2021). The second concern regards the evaluation of ChatGPT's performance due to the closed nature of this model and its continuous updates via Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) (see Section 3.1.1). If the first bias has social implications, the second may compromise a fair evaluation of the model. This is because it is not possible to assess if ChatGPT has been contaminated, seeing during one of its training phases evaluation datasets. This problem has been demonstrated in other recent LLMs (Carlini et al., 2021; Dodge et al., 2021).

To deepen these concerns, we propose a new evaluation based on specialist knowledge. We developed DynaKnowledge, a new benchmark for QA based on information extracted from Wikipedia and related to non-popular entities in the music domain. In this way, we can test how reliable ChatGPT is on very specific questions and a dataset that has never been published. However, it is im-

LREC-COLING 2024, pages 10172–10185

portant to note that the questions in our benchmark are not impossible to answer. This is because we extracted questions and answers from the English-language Wikipedia, one of the most popular sources in NLP. We developed this evaluation inspired by the idea that LLMs have difficulties dealing with low-frequency entities and that if prompted with them, they are likely to *hallucinate*, producing fluent but unrealistic text. We compared the performance of ChatGPT with a knowledge extraction pipeline that uses explicit knowledge representation in the form of Abstract Meaning Representation (Banarescu et al., 2013, AMR) graphs to answer the questions. This explicit knowledge is structured as a Knowledge Graph (Hogan et al., 2021, KG) and thus can be queried. Such a structure also provides the source from which the answer was extracted, a desirable feature also for LLMs but not directly derivable from them, given that they are based on latent and not explicit representations.

The contributions of this paper can be listed as follows:

- a new dynamic benchmark for open-format QA models;
- a new 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.

DynaKnowledge and the evaluation results can be found at `https://github.com/polifonia-project/llms-vs-specialised-knowledge`.

2. Related Work

In this section, we first introduce recent works on the evaluation of ChatGPT, and then we present recent works about the use of AMR for knowledge extraction and QA.

2.1. ChatGPT evaluation

Laskar et al. (2023) proposed a systematic evaluation of ChatGPT in NLP tasks covering 140 tasks, including QA, text summarization, code generation, and commonsense reasoning, and analyzing 255K responses. The results of this large-scale evaluation show that ChatGPT can perform a wide variety of tasks with impressive performance. However, it is still far from achieving good performances in some of them. A domain-specific evaluation of ChatGPT was proposed by Jahan et al. (2023). This evaluation centred on the biomedical domain and covered tasks such as relation extraction, document classification, QA, and summarization. The authors of this work found that ChatGPT in a zero-shot setting can outperform fine-tuned generative models, such as BioGPT (Luo et al., 2022) and BioBART (Yuan et al., 2022), only when the training set for the task is small. Otherwise, if the training set for fine-tuning is large, specialized models outperform ChatGPT by a large margin.

A different evaluation aimed at checking how the predictions of ChatGPT change over time was presented by Aiyappa et al. (2023). The authors of this work used stance detection as a case in point and followed the experimental setting proposed in previous work by Zhang et al. (2022). In this work publicly available data from SemEval 2016 Task 6 (Mohammad et al., 2016, 2017) and P-stance (Li et al., 2021) were used. The authors prompted ChatGPT using, for each sentence in the dataset, the following template: *what's the attitude of the sentence:* INPUT SENTENCE *select from "favor, against or neutral"*. The evaluation shows big improvements comparing results from different time periods, highlighting the difficulty of evaluating this model that is continuously updated with unknown data. The authors concluded that a fair evaluation of this model should be novel in order to prevent test data leakage.

A common view in the literature concerns the closed nature of this model and the high variability of the results, which change substantially depending on apparently uncorrelated factors, such as the moment in which the model was used or a slight modification of the prompt. In such conditions, the parameters of the model, such as the temperature that controls the *creativity of the model*, are difficult to tune and adapt to the task at hand.

2.2. AMR for Knowledge Extraction and Question Answering

Graph-based semantic parsing has gained attention due to the potential of general-purpose representations, such as AMR. Text-to-AMR transduction based on neural machine translation and sequence-to-sequence (*seq2seq*) models achieved high performances (around 85% accuracy) both in English (Bevilacqua et al., 2021; Zhou et al., 2021), multilingual parsing (Blloshmi et al., 2020; Cai et al., 2021) and multi-formalisms scenarios (Procopio et al., 2021).

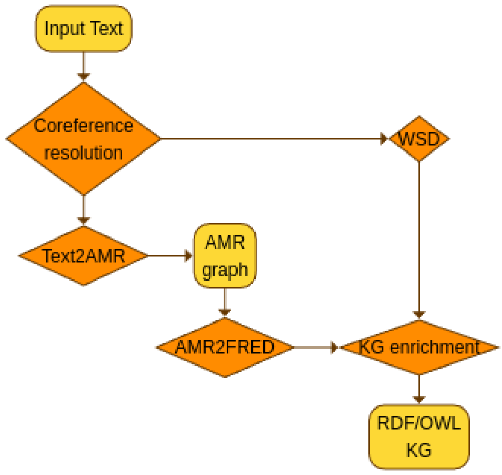
Regarding information extraction end-tasks, Rao et al. (2017) leveraged AMR semantic parsing for extracting information (molecular events/interactions) from biomedical documents, achieving promising results. With a similar objective (extracting fine-grained information from the biomedical scientific literature), Zhang et al. (2021) employed AMR to obtain an in-depth semantic structure of sentences and further enriched the resulting graphs with domain-specific information from external KBs through edge-conditioned graph attention network (GAT, (Velickovic et al., 2018)). The fusion of the semantic structures uncovered by the AMR parsing and the highly specialistic knowledge integrated from a domain-focused KB brought tangible advantages for achieving high performances in the end-task goal.

Kapanipathi et al. (2021) satisfactorily employed AMR as an intermediate logic for the semantic parsing of natural language to develop a Knowledge Base Question Answering (KBQA) neuro-symbolic system. Similarly, Zhang and Ji (2021) proved the advantages of employing the rich semantic representation offered by pre-trained AMR parsers to extract entities, relations, and events from unstructured sentences and encapsulate them into Information Networks through an encoder-decoder architecture. In Lim et al. (2020), Commonsense QA is pursued considering the importance of representing the question's meaning to predict the correct answer. The semantic parsing of the question is obtained by expanding an AMR graph with an external KG, which contains commonsense information (ConceptNet, (Speer et al., 2017)). The resulting graph is subsequently pruned to retain only the most informative knowledge to accomplish the task, then used to analyse the reasoning path and derive the correct answer. Kapanipathi et al. (2021) proposes a Neuro-Symbolic Question Answering (NSQA) system, in which AMR parsing is used to understand questions, and the resulting AMR graph is mapped into query graphs (SPARQL queries) aligned with a KB utilising deterministic mappings. Bornea et al. (2021) considers those deterministic mappings as prone to coverage and granularity mismatch and proposes a framework in which a transition-based semantic parser which integrates a BART-based model learns to transpile AMR into the SPARQL training language. Deng et al. (2022) employs AMR-based modules (AMR parsing, AMR graph segmentation and AMR-To-Text generation) to better understand complex questions through question decomposition and achieve more interpretable multi-hop QA.

3. Methodology

In this paper, we want to compare the performance of explicit and latent knowledge models on the QA task. To this end, we first collected a new benchmark for the task and then selected two representative models for latent and explicit knowledge representation. The latent representation model is ChatGPT, and the explicit knowledge model is a Knowledge Extraction pipeline whose core component is (Gangemi et al., 2023, Text2AMR2FRED), a framework based on Abstract Meaning Representation (AMR) (Banarescu et al., 2013) that produces graph-based semantic representations of the textual evidence we employed for the QA task. In the following sections, we describe the models and then the dataset.

Figure 1: The Knowledge Extraction pipeline schema.



3.1. Models

3.1.1. ChatGPT

As far as it is known, ChatGPT is trained in a three-step process (OpenAI, 2022). First, an initial LLM based on GPT 3/3.5 (Ouyang et al., 2022) is fine-tuned on a dataset created by asking human annotators to write what is the desired output to prompts submitted to the OpenAI API. The next phase consists of sampling a set of prompts from a larger collection of prompts submitted to the OpenAI API. For each of them, the LLM produces multiple responses that human annotators then rank. With this information starts the second training phase in which a reward model is learned on the response-ranking task using the reinforcement learning from human feedback (RLHF) mechanism (Ouyang et al., 2022). This step keeps the LLM frozen and solely trains the RM. In the last training step, the LLM generates responses to a new set of prompts that were not included in the previous steps. In this phase, the now-frozen reward model is used as a reward function, and the LLM is further fine-tuned to maximize this reward using the Proximal Policy Optimization algorithm (Schulman et al., 2017).

Table 1: Example of data collection for the named entity corresponding to the Wikipedia page `Teresina_Brambilla`, `Giuseppina_Ronzi_de_Begnis` and `Wolfgang_Amadeus_Mozart`. ID column reports the sample's unique identifier in DynaKnowledge to facilitate the reader in retrieving the sample in the benchmark.

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.

3.1.2. The Knowledge Extraction pipeline

The knowledge extraction pipeline that we used in this paper relies on Text2AMR2FRED, which exploits AMR parsing to extract knowledge from unstructured text. AMR formalism is grounded in PropBank's *Frames* (Palmer et al., 2005), which constitute the core lexicon of this resource. Frames consist of predicate-argument structures named *rolesets*. AMR can be expressed by using the PENMAN serialisation format (Matthiessen and Bateman, 1991), a notation convention that enables encoding the semantic dependencies of directed and rooted graphs such as AMR graphs. In our framework, depicted in Figure 1, AMR graphs serve as an event-centric representation of an input text, well-suited for retrieving the essential elements of a situation described in one or multiple sentences. As we can see for example in 2, the Argo of the PropBank predicate sing-01 encodes information about the singer (an *agent* in an act of singing). The advantage of AMR representation is that it detaches from syntax variability and word forms, providing the same graph for sentences conveying almost the same meaning with a different syntactic or lexical realisation. Our Knowledge Extraction pipeline extracts information from the text and stores it in AMR-based KGs through its core component, Text2AMR2FRED.

Implementation details. In our Knowledge Extraction pipeline, we implement techniques for minimising the loss of information in the source text, such as coreference resolution, i.e., the task of clustering spans of text (*mentions*) that correspond to the same single entity (*referent*). For our experiments, we used the model proposed by Clark and Manning (2016) and applied referent substitutions only to a subset of pronominal mentions, the third-person personal pronouns. We evaluated the substitutions performed, achieving a 92% accuracy on a set of 52 third-person pronominal mentions occurring in a sample of sentences extrapolated from Wikipedia.

The text2AMR parsing module of our Knowledge Extraction pipeline relies on a neural semantic parser, i.e., (Bevilacqua et al., 2021, SPRING), which allows us to perform the AMR2text task with the same model. This model exploits BLINK (Wu et al., 2020) to link named entities to their unique entry in Wikipedia (*wikification*).

As a last step, our Knowledge Extraction pipeline includes (Meloni et al., 2017, AMR2FRED), a tool that transforms AMR graphs into RDF/OWL KGs that follow (Gangemi et al., 2017, FRED) machine reader's knowledge representation patterns. The resulting KGs are further enriched with knowledge from external Knowledge Bases (KBs) through Framester (Gangemi et al., 2016), a semantic resources hub. An example of knowledge used to enrich the resulting Knowledge Graphs is Word Sense Disambiguation (WSD) information, which is obtained by submitting the original sentence to (Bevilacqua and Navigli, 2020, EWISER) and used to associate WordNet's synsets from the RDF version of WordNet[1], included in Framester, to AMR nodes missing links to any external source.

Text2AMR2FRED APIs[2] enable tools such as the Machine Reading suite[3] to generate RDF *named graphs*, which allows reporting information about the source of the input text analysed. Text2AMR2FRED is also released to the public via a user-friendly web app[4].

[1] https://www.w3.org/TR/wordnet-rdf/
[2] http://framester.istc.cnr.it/txt-amr-fred/api/docs
[3] https://github.com/polifonia-project/machine-reading
[4] https://arco.istc.cnr.it/txt-amr-fred/

3.2. DynaKnowledge: a new Dynamic Benchmark for QA

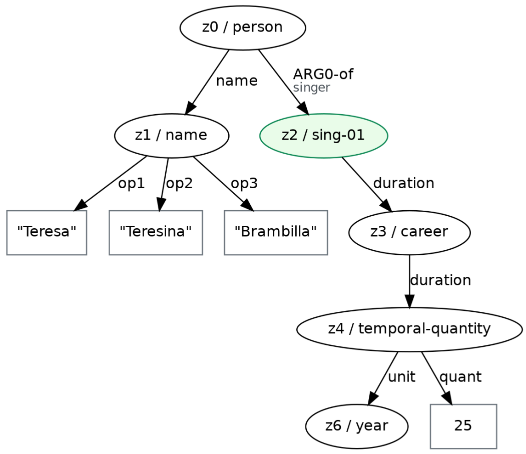
With this paper, we introduce DynaKnowledge, a new manually curated dataset for open-form QA made of questions based on the Wikipedia pages of historical entities picked from a corpus of music historical periodicals[5], whose publication dates range from 1823 to 1900. The dataset comprises 82 samples (question, answer, textual evidence tuples). The textual evidence element consists of the Wikipedia page sentences containing the answer to the question. To ensure that our benchmark's questions are not impossible to answer, we included only questions that were unambiguously answerable by reading a passage from English-language Wikipedia. In fact, English-language Wikipedia is commonly included in LLMs training sets, for example, in the GPT-3 training set (Brown et al., 2020).

Given the nature of the corpus, the selected historical characters are domain-specific (music) and known to be active or alive before the periodicals' publication dates range. We maintained a 50-50 gender distribution between the selected named entities[6]. As a result, half of the questions focus on male historical characters and the other half on female historical characters. DynaKnowledge is intended to be dynamic and updated over time with new samples always related to low-frequency entities.

3.2.1. Data collection

Data collection was performed by Foreign Languages and Literature undergraduate students who received *ad hoc* training as part of their bachelor's degree curricular internship. They recorded the data in a spreadsheet following the format as in the sample reported in Table 1. On a more nuanced level, the analysts (i) selected a historical character (technically, a named entity of type PERSON) occurring in a corpus of music historical periodicals; (ii) accessed the chosen historical character's Wikipedia page and selected one or more sentences reporting a fact relating to the entity; (iii) formulated a question related to the historical character, answerable by the selected Wikipedia page sentences.

Figure 2: AMR subgraph containing the answer to the question *"How long was Teresina Brambilla career as a musician?"*.



For each sample, we report the following information:

1. ID: the sample's unique identifier;
2. Date: the date, or range of dates, on which the sample was collected;
3. Wikipedia page: the Wikipedia page of the sample's named entity;
4. Gender: the gender of the sample's named entity;
5. Question: the question about a fact concerning the sample's named entity;
6. Prompt: the prompt submitted to ChatGPT, containing the question;
7. Answer: the ground truth answer;
8. Provenance: the textual evidence, i.e. the sentence(s) from the entity's Wikipedia page containing the answer to the question.

The data collection method that we followed for DynaKnowledge is designed to support an ever-evolving benchmark continuously enriched with new examples that focus on less common entities.

4. Experimental setup

We tested GPT-3.5 (Default) model, accessed via its free online version (ChatGPT[7] and our Knowledge Extraction pipeline (cfr. 3.1.2). We designed

[5] This corpus is the *Periodicals* module of the broader Polifonia Textual Corpus (https://github.com/polifonia-project/Polifonia-Corpus). As we cannot release the module's data for copyright reasons, we release at https://zenodo.org/records/6671912 its metadata containing title, year, and issue number information, allowing for its complete reconstruction.

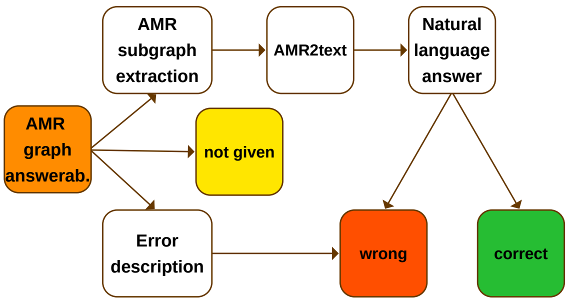
[6] In our study, we restrict to binary gender categories, which, although not reflecting real-world diversity, let us move the first steps towards the definition of our method. We consider a *female* person named entity a historical character that is referred to, in its Wikipedia page, with the pronoun *she*, a *male* person named entity a historical figure who is referred to, in its Wikipedia page, with the pronoun *he*.

[7] https://chat.openai.com/

Table 2: Evaluation - ChatGPT's answer assessment. ID column reports the sample's unique identifier in DynaKnowledge.

ID	Evaluator's Prompt	ChatGPT's answer	Assessment
18	How long was Teresina Brambilla's career as a musician? Give a straight 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. Therefore, I cannot provide information about the length of her musical career. It's possible that she gained recognition after my last update or that she is not a prominent figure in the music world.	Not given
41	Which female colleague did Giuseppina Ronzi de Begnis argue with during the rehearsals of Maria Stuarda? Give a straight answer.	Giuseppina Ronzi de Begnis argued with Maria Malibran during the rehearsals of Maria Stuarda.	No
5	How old was Wolfgang Amadeus Mozart when he started to compose? Give a straight answer.	Wolfgang Amadeus Mozart started composing at the age of 5.	Yes

Figure 3: Process followed by the analysts to record the AMR graph's *answerability* assessment. If *answerability* is satisfied, the analysts extract the subgraph containing the answer to the question that is then transformed into text via AMR2text to confirm the AMR graph's *answerability*. If not, they explain the error observed in the graph.



and executed the experiments from August 4th to August 22nd, 2023. At that time, it was not possible for a user to fine-tune GPT-3.5 (Default) model. Therefore, we evaluated it in a zero-shot setting. We tested GPT-3.5 (Default) and our Knowledge Extraction pipeline on the same questions formulated and collected in the data collection phase (cfr. 3.2.1). We compared the free text answers to the questions returned by ChatGPT and the AMR graphs produced by parsing, with our Knowledge Extraction pipeline, the Wikipedia sentences containing the answer to the question.

4.1. Evaluation

To perform the evaluation, the analysts had to compare ChatGPT's answers to the questions and AMR graph's *answerability*. The evaluation was performed by the same interns who conducted the data collection, with the support of the authors of this paper for the AMR graph *answerability* assessment.

4.1.1. ChatGPT answer

To assess ChatGPT's answer, the analysts had to submit the questions as prompts to the free online version of ChatGPT and manually assess the answer returned by ChatGPT. They recorded the results by adding the following information to the dataset described at 3.2:

1. ChatGPT answer: ChatGPT's answer;
2. ChatGPT answer assessment (yes/no/not given): manual evaluation of whether the ChatGPT response is correct (*"yes"*), incorrect (*"no"*), or the model does not give any answers (*"not given"*).

Given the reduced size of our sample, we manually compared ChatGPT's and ground truth answers to ensure the high quality of the assessment. Such a choice was also made in light of the awareness raised by Chen et al. (2019) and Roberts et al. (2020) on how QA metrics based on tokens overlap can hinder the evaluation of questions that require more complex and abstractive answers.

4.1.2. AMR graph

We manually checked the AMR graphs to assess AMR graphs' *answerability* to guarantee the highest assessment quality, as stressed out by Ettinger et al. (2023). The process is depicted in Figure 3. The analysts transformed the Wikipedia sentences containing the answer to the question into an AMR graph via Text2AMR2FRED and then assessed the *answerability* of the AMR graph. We define *answerability* as the ability of the AMR graph to encode the information needed to answer a question. It is satisfied when the AMR graph contains a subgraph that correctly captures the information needed to answer a question. The subgraph must report the semantically correct PropBank predicates, their core roles (arguments), and non-core roles (relations) structure. It must be possible to translate the subgraph back to text using an AMR2text model and obtain a natural language sentence containing an answer to the question. They recorded the AMR graphs' answerability assessment results by adding the following information to the dataset described at 3.2:

1. AMR graph (PENMAN): the AMR graph corresponding to the Provenance (cfr. 3.2.1);

2. AMR GRAPH'S ANSWERABILITY (YES/NO/NOT GIVEN): manual assessment of the *answerability* of the AMR graph;

3. IF SO, WHICH IS THE SUB-GRAPH CONTAINING THE ANSWER?: the AMR sub-graph containing the answer to the question;

4. AMR GRAPH'S ERROR DESCRIPTION: if the value reported in DOES THE AMR GRAPH CONTAIN AN ANSWER TO THE QUESTION? is "no", an explanation of the error observed.

4.1.3. Metrics

We started from the assumption that finding the most suitable metrics for QA is a difficult task and that a metric such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004) or METEOR (Banerjee and Lavie, 2005) can alter the evaluation of free-form QA (Chen et al., 2019). With this in mind, as we presented in the previous section, we manually checked the answers of the two models under analysis to ensure that the results were not biased. With the answers manually checked and divided into the canonical *correct*, *wrong*, and *not given*, we are able to compute standard precision (P), recall (R) and F1 (F1) as:

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

and

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

5. Results and Analysis

Table 3 compares ChatGPT answers and the AMR graphs' answerability assessment. The table shows that the AMR graphs effectively capture the answers to the questions 50% of the times, while ChatGPT's answer is correct 38% of the times (precision). We can also see that the recall of ChatGPT is higher than precision. ChatGPT, in fact, can give an answer that indicates that it is not able to provide information. On the other hand, this is not possible with AMR parsing. We considered a question prompted to AMR as *not given* only if the parser cannot produce a graph. This happened just one time in our experiments.

In the example with ID `18` reported in Table 2, ChatGPT's answer has been assessed as *not given*. In fact, ChatGPT replies to the prompt stating that, to the best of its knowledge, there is no known musician named *Teresina Brambilla*, and for this reason, it cannot answer the question asked (*"How long was Teresina Brambilla's career as a musician?"*). But, as recorded in Table 1, the sentence from Wikipedia containing the answer (PROVENANCE) collected for this sample states that the correct answer for the sample's question is *25 years*. In the example with ID `41` reported in Table 2, ChatGPT's answer assessment is *no*, meaning that the answer given by ChatGPT is not correct. In fact, as recorded in Table 1, the sentence from Wikipedia containing the answer (PROVENANCE) collected for this sample reports that the correct answer for the sample's question (*"Which female colleague did Giuseppina Ronzi de Begnis argue with during the rehearsals of Maria Stuarda?"*) is *Anna Del Sere*, while ChatGPT mentions an argument between Giuseppina Ronzi and Maria Malibran. In the example with ID `5` reported in Table 2, ChatGPT's answer assessment is *yes*. ChatGPT's answer correctness is verifiable by comparing ChatGPT's answer in Table 2 with the analyst's answer reported in Table 1: both contain the same piece of knowledge, namely that Mozart started composing at the age of five.

Table 3: ChatGPT's answers and AMR graphs' answerability assessment reported as precision (P), recall (R), f1-measure (F1) dividing the questions according to the gender of the named entity involved, female (F) and male (M).

NE's gender	ChatGPT P	ChatGPT R	ChatGPT F1	AMR P	AMR R	AMR 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**

Figure 2 reports the AMR subgraph containing the answer to the question formulated for the sample with ID `18` (cfr. Table 1). The subgraph, extrapolated from the AMR graph resulting from the text2AMR parsing of the PROVENANCE (sentence from Wikipedia containing the answer) of the same sample is considered able to answer the question. It has a *root node*, Z0/SOPRANO, that is linked to the *instance node* Z1/PERSON by the *non-core role* :DOMAIN. This triple correctly reflects the *"named-entity-is-noun"* semantics. The *instance node* Z1/PERSON is linked to the *instance node* Z2/NAME by the *non-core role* :NAME. This triple correctly states that the *instance node* Z1/PERSON has a *name*, therefore is a *named entity*. It branches out in three leaves nodes linked by the relation :OP, which reports the three tokens of the sentence that make up the named entity mention (*Teresa*, *"Teresina"*, *Brambilla*). The *instance node* Z1/PERSON is also linked to the PropBank *predicate* node Z8/SING-01 by the *inverse core role* :ARG0-OF. This triple correctly captures the semantics of relative clauses. SING-01 is also linked to the node Z14/CAREER by the *non-core role* :DURATION, which is itself linked through the same relation to the *instance node z15/temporal-quantity*, which captures the semantics of a specific type of quantity. It

Table 4: The Wikidata identifier (QID) average frequency and standard deviation distribution across the named entities of our sample, split per gender and ChatGPT answers'/AMR graphs' answerability assessment

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

Table 5: Correlation between the gender/popularity of the named entities of our sample and ChatGPT's answers/AMR graphs' answerability assessment reported as Spearman's Rho coefficient (ρ_s).

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

branches out into the two tokens of the sentence that make up the temporal quantity mention by its arguments :UNIT (*z16/year*) and :QUANT (*25*).

The semantic structure of this subgraph encloses all the elements required to answer the question asked, as can also be verified by transforming the AMR graph back to text employing the same model used for Text2AMR, i.e., SPRING. The resulting natural language sentence *Teresa Teresina Brambilla is a soprano who has sung for a 25-year career* answers correctly the question asked (cfr. the column `Analyst's Answer` in Table 2).

In the AMR graph resulting from the text2AMR transformation of the sentence from Wikipedia containing the answer (PROVENANCE) of the sample with ID 41, the named entity *Maria Stuarda* is incorrectly classified as a PERSON instead of as a WORK-OF-ART:

```
(z16 / person
    :name (z17 / name
    :op1 "Maria"
    :op2 "Stuarda"))
```

The named entity classification reported in the example above compromises the correctness of the resulting AMR graph. The AMR graph *answerability* criteria is not satisfied.

5.1. Issues

While assessing AMR graphs' *answerability*, we encountered two main issues.

Coreference resolution Some of the sentences collected from Wikipedia pages (cfr. 3.2.1) contain only mentions of the entity and not the entity's referent. Mentions can be *personal pronouns* (as in the sentence *"He composed from a very young age, first studying with his uncle Giovanni Mazzetti and later with Luigi Caruso in Perugia"*), *possessive pronouns* (as in the sentence *"His parents were John Reeves, a musician of Yorkshire origin, and his wife, Rosina"*), *names* (i.e., proper nouns in their shortened forms, as in the sentence *"Barnby was born at York, as a son of Thomas Barnby, who was an organist"*). As reported in paragraph 3.1.2, in our Knowledge Extraction pipeline, we apply coreference resolution to only a subset of pronominal mentions (the third-person personal pronouns). As a consequence, AMR subgraphs extrapolated from sentences in which the only mention of the entity is a possessive pronoun or a surname are never assessed as correct because, at present, we do not manage them in our Knowledge Extraction framework. Instead, we considered AMR subgraphs extrapolated from sentences in which the only reference to the entity was a third-person personal pronoun as correct when they corresponded to our *answerability* criteria.

Wikification In assessing AMR graphs *answerability*, we do not assess the correctness of the entity linking information. Still, we focus on the correct recognition of the entity mention and on its correct classification to a pre-defined entity type[8]. This is because, in our particular case, we are using Wikipedia documents, from which it is possible to extract unambiguous entity references by disposing of entity linking altogether.

5.2. Popularity and gender effect

Looking at the results in Table 3 in a more granular way, we notice that ChatGPT's answers tend to be wrong more often when questions regard female historical characters: ChatGPT's answer is correct 22% of the times when prompted with questions regarding female named entities, and 54% of

[8]Entity types list (`https://www.isi.edu/~ulf/amr/lib/ne-types.html`) and criteria for their assignation are defined in AMR annotation guidance instructions.

the times when prompted with questions regarding male named entities. These observations might corroborate the hypotheses of gender bias in LLMs (Cheng et al., 2023) and in KBs such as Wikipedia (Stranisci et al., 2023).

Also, we wanted to verify whether the named entities' popularity influenced the results, as other research suggests (Chen et al., 2021) (Kandpal et al., 2023). We define *popularity* as each named entity's Wikidata identifier (QID) frequency of occurrence as an internal link in Wikipedia[9]. We added the following two pieces of information to the benchmark described in 3.2:

1. QID: The QID corresponding to the sample's named entity Wikipedia page;
2. POPULARITY (QID_FREQUENCY): number of times an internal link in Wikipedia can be mapped to a QID.

As we can see in Table 4, male named entities have, in our dataset, an average QID frequency of 545 against 36 of the female named entities. Suppose we focus on the sub-sample of cases in which ChatGPT returns the highest number of correct answers: those cases correspond to male historical characters with the highest QID frequency average across the whole sample (869).

To strengthen our hypothesis, we calculate the correlation between ChatGPT's answers' assessment and named entities' popularity using Spearman's Rho coefficient (ρ_s). As we can see in Table 5, the result obtained for popularity ($\rho_s = 0.48$) show a positive correlation. We repeat the experiment for AMR graphs' answerability, and the results obtained ($\rho_s = -0.07$) demonstrate no correlation. When considering gender as a feature, we obtain the same results: a positive correlation with ChatGPT's answers' assessment ($\rho_s = 0.33$), no correlation ($\rho_s = -0.02$) with AMR graphs' answerability.

Notably, the AMR graphs' answerability is apparently robust to any major bias toward the gender or popularity of the entities involved. With regard to ChatGPT's answers' assessment, the popularity bias seems to amplify the gender bias: we eventually obtain wrong answers more often when our questions regard female named entities, which in our sample also have the lowest average popularity. This confirms our hypothesis that a simpler knowledge extraction pipeline is more robust to variation in named entities' features, such as popularity and gender, than an LLM such as ChatGPT.

6. Future work

In future work, we plan to expand our benchmark by adding more samples. The way in which we want to develop the benchmark is dynamic, aligning with current trends in benchmarking NLP models (Kiela et al., 2021). We believe that it is necessary since models like ChatGPT are continuously updated, and an answer from these models can vary over time or with slight changes to the prompt. This will allow us to examine further the interplay between gender and popularity biases and to deepen the investigation regarding the difficulties that LLMs demonstrate when dealing with long-tail knowledge (Kandpal et al., 2023).

We also plan to transform the AMR graphs output by our Knowledge Extraction pipeline into Resource Description Framework (RDF) KGs, as Text2AMR2FRED allows. This will enable us to automate the evaluation through structured interrogations and develop an automatic evaluation metric.

Eventually, we aim to combine the capabilities of both the latent and explicit knowledge representation models in a single framework in line with retrieval-augmented generative models (Lewis et al., 2020) and testing the resulting model on the same benchmark. Following Mallen et al. (2023), we want to assess whether using the explicit knowledge produced by our knowledge extraction pipeline to apply constraints on LLM's output will enhance the results and reduce hallucination, mitigating gender and popularity bias on domain-specific and long-tail entities as the ones in DYNAKNOWLEDGE.

7. Conclusion

In this paper, we conducted an evaluation of explicit and latent knowledge representation models. This evaluation was possible thanks to the development and release of DYNAKNOWLEDGE, a new benchmark for the open-form QA task. The samples included in DYNAKNOWLEDGE revolve around named entities of type person extrapolated from historical periodicals. The benchmark's questions are all answerable by reading a passage from Wikipedia. This first release serves as a starting point for developing a dynamic benchmark that will be updated over time to tackle the continuous improvements that current LLMs are having. The results of our experiments show that our benchmark is particularly challenging. In particular, we found that ChatGPT struggles to answer questions related to less popular entities. This result was accompanied by the assessment of the higher performances of a simple knowledge extraction pipeline. Such a pipeline demonstrated to be more robust on the variation of named entities' features such as gender and popularity.

[9]We used the `enwiki-20220120` dump.

8. Limitations

In this paper, we want to provide a feasibility study for developing a fully automatic model for open-domain QA founded on AMR-based KGs. On the one hand, we wanted to test whether our knowledge extraction pipeline correctly encoded the information required to answer the questions, storing it in KGs that can be reliably queried to obtain the answers. On the other hand, we shed light on the sub-optimal performance of ChatGPT when elicited to output factual knowledge about less popular named entities. We concentrated on ChatGPT among all the LLMs because of its widespread popularity and commoditization. The capability of our knowledge extraction pipeline to provide explicit knowledge that can be automatically queried to obtain reliable answers or that can augment ChatGPT performance on open domain QA over long-tail entities should be tested in a fully automatised scenario, in which subgraphs satisfying the answerability condition are automatically retrieved.

9. Acknowledgements

Polifonia

The authors gratefully acknowledge the support of the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004746.

Funded by the Horizon 2020 Framework Programme of the European Union

10. Bibliographical References

Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-yeol Ahn. 2023. Can we trust the evaluation on ChatGPT? In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 47–54, Toronto, Canada. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking". In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event, Toronto, Canada. Association for Computing Machinery.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12564–12573.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Rexhina Blloshmi, Rocco Tripodi, and Roberto Navigli. 2020. XL-AMR: Enabling Cross-Lingual AMR Parsing with Transfer Learning Techniques. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2487–2500, Online. Association for Computational Linguistics.

Mihaela Bornea, Ramón Fernandez Astudillo, Tahira Naseem, Nandana Mihindukulasooriya, Ibrahim Abdelaziz, Pavan Kapanipathi, Radu Florian, and Salim Roukos. 2021. Learning to transpile AMR into SPARQL. ArXiv preprint arXiv:2112.07877.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language

Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, Online. Curran Associates, Inc.

Deng Cai, Xin Li, Jackie Chun-Sing Ho, Lidong Bing, and Wai Lam. 2021. Multilingual AMR parsing with noisy knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2778–2789, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *Proceedings of the 30th USENIX Security Symposium*, pages 2633–2650, online. USENIX Association.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4472–4485, Online. Association for Computational Linguistics.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532, Toronto, Canada. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Zhenyun Deng, Yonghua Zhu, Yang Chen, Michael Witbrock, and Patricia Riddle. 2022. Interpretable AMR-Based Question Decomposition for Multi-hop Question Answering. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4093–4099, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, pages 1286–1305, Virtual Event, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Allyson Ettinger, Jena Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. "you are an expert linguistic annotator": Limits of LLMs as analyzers of Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8250–8263, Singapore. Association for Computational Linguistics.

Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti, and Diego Reforgiato Recupero. 2016. Framester: A Wide Coverage Linguistic Linked Data Hub. In *Knowledge Engineering and Knowledge Management*, pages 239–254, Bologna, Italy. Springer International Publishing.

Aldo Gangemi, Arianna Graciotti, Antonello Meloni, Andrea Nuzzolese, Valentina Presutti, Diego Reforgiato Recupero, Alessandro Russo, and Rocco Tripodi. 2023. Text2AMR2FRED, a tool for transforming text into RDF/OWL Knowledge Graphs via Abstract Meaning Representation. In *22nd International Semantic Web Conference*, Athens, Greece. CEUR Workshop Proceedings.

Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic Web Machine Reading with FRED. *Semantic Web*, 8(6):873–893.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938, Online. Proceedings of Machine Learning Research.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. *Knowledge Graphs*. Synthesis Lectures on Data, Semantics, and Knowledge. Morgan & Claypool Publishers.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 326–336, Toronto, Canada. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12).

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. Proceedings of Machine Learning Research.

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramón Fernandez Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue, Dinesh Garg, Alfio Gliozzo, Sairam Gurajada, Hima Karanam, Naweed Khan, Dinesh Khandelwal, Young-Suk Lee, Yunyao Li, Francois Luus, Ndivhuwo Makondo, Nandana Mihindukulasooriya, Tahira Naseem, Sumit Neelam, Lucian Popa, Revanth Gangi Reddy, Ryan Riegel, Gaetano Rossiello, Udit Sharma, G P Shrivatsa Bhargav, and Mo Yu. 2021. Leveraging Abstract Meaning Representation for Knowledge Base Question Answering. In *Findings of the Association for Computational Linguistics: Association for Computational Linguistics-IJCNLP 2021*, pages 3884–3894, Online. Association for Computational Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: Association for Computational Linguistics 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, Online. Curran Associates, Inc.

Jungwoo Lim, Dongsuk Oh, Yoonna Jang, Kisu Yang, and Heuiseok Lim. 2020. I Know What You Asked: Graph Path Learning using AMR for Commonsense Reasoning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2459–2471, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9):195:1–195:35.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Christian M. I. M. Matthiessen and John A. Bateman. 1991. *Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese*. Pinter Publishers.

Antonello Meloni, Diego Reforgiato Recupero, and Aldo Gangemi. 2017. Amr2fred, a tool for translating abstract meaning representation to motif-based linguistic knowledge graphs. In *The Semantic Web: ESWC 2017 Satellite Events*, pages 43–47, Cham. Springer International Publishing.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, Online and New Orleans, LA, USA. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Luigi Procopio, Rocco Tripodi, and Roberto Navigli. 2021. SGL: Speaking the Graph Languages of Semantic Parsing via Multilingual Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 325–337, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017. Biomedical Event Extraction using Abstract Meaning Representation. In *BioNLP 2017*, pages 126–135, Vancouver, Canada. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. ArXiv preprint arXiv:1707.06347.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Marco Antonio Stranisci, Rossana Damiano, Enrico Mensa, Viviana Patti, Daniele Radicioni, and Tommaso Caselli. 2023. WikiBio: a semantic resource for the intersectional analysis of biographical events. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12370–12384, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, Long Beach, CA, USA. Curran Associates Inc.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. ArXiv preprint arXiv:1710.10903.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? ArXiv preprint arXiv:2212.14548.

Zixuan Zhang and Heng Ji. 2021. Abstract Meaning Representation guided graph encoding and decoding for joint information extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–49, Online. Association for Computational Linguistics.

Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. 2021. Fine-grained Information Extraction from Biomedical Literature based on Knowledge-enriched Abstract Meaning Representation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6261–6270, Online. Association for Computational Linguistics.

Jiawei Zhou, Tahira Naseem, Ramón Fernandez Astudillo, Young-Suk Lee, Radu Florian, and Salim Roukos. 2021. Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based AMR parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

11. Language Resource References

Li, Yingjie and Sosea, Tiberiu and Sawant, Aditya and Nair, Ajith Jayaraman and Inkpen, Diana and Caragea, Cornelia. 2021. *P-Stance: A Large Dataset for Stance Detection in Political Domain*. Association for Computational Linguistics.

Mohammad, Saif and Kiritchenko, Svetlana and Sobhani, Parinaz and Zhu, Xiaodan and Cherry, Colin. 2016. *SemEval-2016 Task 6: Detecting Stance in Tweets*. Association for Computational Linguistics.

Mohammad, Saif M. and Sobhani, Parinaz and Kiritchenko, Svetlana. 2017. *Stance and Sentiment in Tweets*. Association for Computing Machinery.

Palmer, Martha and Gildea, Daniel and Kingsbury, Paul. 2005. *The Proposition Bank: An Annotated Corpus of Semantic Roles*. MIT Press.

Speer, Robyn and Chin, Joshua and Havasi, Catherine. 2017. *ConceptNet 5.5: an open multilingual graph of general knowledge*. AAAI Press, AAAI'17.