# CR-LLM: A Dataset and Optimization for Concept Reasoning of Large Language Models

**Nianqi Li**[1], **Jingping Liu**[2], **Sihang Jiang**[1*], **Haiyun Jiang**[1], **Yanghua Xiao**[1*], **Jiaqing Liang**[3], **Zujie Liang**[4], **Feng Wei**[4], **Jinglei Chen**[4], **Zhenghong Hao**[4], **Bing Han**[4]
[1]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
[2]School of Information Science and Engineering, East China University of Science and Technology
[3]School of Data Science, Fudan University [4]MYbank, Ant Group
nqli23@m.fudan.edu.cn, shawyh@fudan.edu.cn, tedsihangjiang@gmail.com

## Abstract

Concept reasoning is an important capability for models to understand the world. However, the existing datasets, such as concept extraction and concept generation, suffer from modeledge leakage and context leakage. To address these limitations, we construct a dataset of concept reasoning for large language models (CR-LLM) with modeledge leakage prevention and context leakage prevention, which consists of 2,167 samples and covers different concept types. In addition, we propose a hybrid reasoning method, consisting of inductive reasoning, deductive reasoning and a controller. This method allows large language models to adaptively select the optimal reasoning method for each input sample. Finally, we conduct extensive experiments on CR-LLM using different models and methods. The results show that existing large language models and reasoning methods perform sub-optimally in the concept reasoning task. In contrast, our proposed method significantly improves the capabilities, achieving a 7% increase in accuracy compared to CoT and demonstrating better granularity. We release CR-LLM and code at https://github.com/Nianqi-Li/Concept-Reasoning-for-LLMs.

## 1 Introduction

Concept reasoning is an important ability for models to understand the world by producing appropriate entity concepts based on contextual information. This ability can enhance the model performance in several downstream tasks, such as entity linking (Yang et al., 2019), text classification (Chen et al., 2019), probing (Peng et al., 2022), event plausibility (Porada et al., 2021), and recommendation systems (Sharma et al., 2017). Although previous studies highlight the reasoning abilities of large language models (LLMs), such as chain-of-thought (Wei et al., 2022) and mathematical reasoning (Gao et al., 2023), there remains a lack of research on concept reasoning. This paper thus focuses on evaluating and improving the concept reasoning abilities of LLMs.

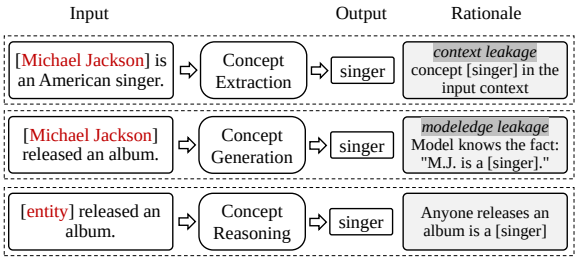


Figure 1: Examples of concept extraction, concept generation, and concept reasoning tasks.

Existing concept reasoning related datasets struggle to provide sufficient support for research on concept reasoning in LLMs. These datasets can be roughly divided into two types: concept extraction datasets and concept generation datasets. Concept extraction datasets takes an entity and its description as input, and extracts concept of the entity from the description as output (Yuan et al., 2021). Concept generation datasets use the same inputs, but the output may not appear in the description (Ling and Weld, 2012). However, as we investigate the rationales provided by LLMs, two types of rationales are detrimental to testing the concept reasoning ability of LLMs: *modeledge leakage* (Han et al., 2021) and *context leakage*. We show some examples in Figure 1. In modeledge leakage, the factual knowledge is stored in LLMs, and the concept can be generated without the entity description. In context leakage, the concept of the given entity is already included in the input, then LLMs can directly extract the answer without reasoning.

To evaluate LLMs' concept reasoning capability without the modeledge leakage and context leakage, we introduce a dataset of concept reasoning for

*Corresponding author.

LLMs (CR-LLM). The input is a context with a masked entity, and the output is the concept for that masked entity in the context, as shown in Figure 1. CR-LLM excludes target entities and concepts in input to prevent modeledge leakage and context leakage.

The evaluation based on CR-LLM shows that existing widely-used LLMs achieve limited performance on concept reasoning. Intuitively, concept reasoning can benefit from the logical reasoning ability of LLMs, of which inductive and deductive reasoning are common forms (Lawson, 2005; Bang et al., 2023), we thus propose a hybrid method that combines inductive and deductive reasoning to stimulate the concept reasoning capability of LLMs. Inductive reasoning involves specific observation of patterns to draw a more general conclusion (Lawson, 2005), which inspires us to retrieve similar samples with the input to conclude the possible concepts. On the contrary, deductive reasoning drives specific conclusions based on more general premises (Lawson, 2005), which asks to retrieve relevant facts to infer. Additionally, we design a controller based on LLM, which allows the model to select the appropriate reasoning methods for different samples.

**Contributions.** The contributions of this paper are as follows:

- We are the first to introduce the concept reasoning task for LLMs, which is based solely on the models' reasoning ability, rather than the extraction ability (context leakage) and the factual knowledge (modeledge leakage) stored in LLMs.
- We create a dataset CR-LLM to evaluate the concept reasoning abilities of LLMs, consisting of 2,167 samples and covering various concept types.
- We propose a hybrid reasoning method based on induction and deduction that allows models to adaptively select the appropriate reasoning method for each sample.
- We conduct extensive experiments on CR-LLM using different models and methods. The results show that our proposed method significantly improves the model performance, achieving a 7% increase in accuracy compared to CoT and demonstrating better granularity.

## 2 Related Work

**Concept Related Research** Previous studies on obtaining concepts can be divided into three categories: concept extraction, entity typing and concept generation.

Concept extraction focuses on extracting entity concepts from text using an extractive approach. Inputs include the target entity and text with the concept, while outputs provide the concept of the target entity. Common approaches encompass pattern matching (Auer et al., 2007), learning-based extraction (Nguyen et al., 2019; Nie et al., 2020; Yuan et al., 2021, 2023), and knowledge-based extraction (Bai et al., 2019; Preum et al., 2020; Qiu et al., 2019). Entity typing aims to classify given entities into predefined concept sets. Datasets and methods in this area vary depending on the granularity. Coarse-grained entity typing methods such as BNN (Weischedel and Brunstein, 2005), fine-grained entity classification datasets such as FIGER (GOLD) (Ling and Weld, 2012; Yosef et al., 2012), and ultra-fine entity typing tasks (Choi et al., 2018; Ding et al., 2021) are examples. A smaller body of research uses pre-trained language models to generate entity concepts from text (Yuan et al., 2022), and most studies use datasets shared with entity typing.

However, all concept extraction samples have context leakage as they only extract concepts in text. Entity typing and concept generation also have context leakage: 21.5% of the text contains the output concept in Ultra-fine (Choi et al., 2018), as do 40.7% of the GT-zh dataset (Lee et al., 2020). Furthermore, due to the exposure of entity names, LLMs can directly answer questions using modeledge. Therefore, their datasets and methods do not align with the requirements of concept reasoning in LLMs.

**Reasoning Related Research** Reasoning is often considered a weakness in language models (Bommasani et al., 2021; Rae et al., 2021; Valmeekam et al., 2022). Recent research suggests that reasoning abilities may emerge in LLMs at a certain scale (Wei et al., 2022). Consequently, stimulating the reasoning abilities of LLMs becomes a major research focus. In 2022, Wei et al. (2022) proposed chain-of-thought prompting, which encourages LLMs to engage in reasoning by articulating intermediate steps. Subsequent researches extend this concept: Wang et al. (2022) introduced chain-of-thought based on self-consistency, replac-

ing greedy decoding with multiple paths, and Yao et al. (2023) introduced tree-of-thought to increase the flexibility of reasoning paths. However, given the complexity and diversity of concept reasoning, these paradigms show mediocre performance in this task. Therefore, there is still ample room for improvement and extension of concept reasoning.

## 3 The CR-LLM Dataset

In this section, we first define the task of concept reasoning. Then, we describe the details of the dataset construction and analysis. Finally, we present the evaluation methodology and provide evaluation results on multiple models.

### 3.1 Problem Formulation

Concept reasoning aims to predict the concept of a masked entity within a given sentence. For instance, when presented with the sentence "[entity] released an album", an optimal model is expected to generate the concept "singer" of the masked entity. However, due to the diversity of concepts, the outputs "person" and "musical figure" are also allowed. Therefore, concept reasoning is an open-ended task.

### 3.2 Dataset Construction

In this section, we introduce CR-LLM, a dataset designed to assess the concept reasoning capabilities of LLMs. Each sample comprises a sentence with a masked entity and its concept for reference, as open-ended tasks don't have a unique answer. To avoid the negative impacts of modeledge and context leakages, we utilize the FIGER (Ling and Weld, 2012), a widely-used entity typing dataset, as the source data and employ modeledge leakage prevention method and context leakage prevention method in the construction of samples.

**Modeledge Leakage Prevention** In the context of the FIGER, there are two types of modeledge leakages. In the first case, the model can accurately predict the concept directly from its stored knowledge if the entity in the context is not masked. For example, taking the sentence "[Michael Jackson] released a new song *Beat It*" as input, it is easy for LLMs to predict "Michael Jackson is a singer" without reasoning. To address this, we mask the entity. In the second type, the model can know the masked entity from the context. For example, LLM can indicate the masked entity is "Michael Jackson" according to the song "*Beat It*", and "Michael Jackson" is a "singer", which returns the first type of leakage. Therefore, we use ChatGPT to predict the masked entity in input and remove samples in which masked entities can be predicted.

| Type | #FGRC | #Num | Example |
|---|---|---|---|
| person | 12 | 546 | doctor, actor, artist |
| organization | 11 | 477 | company, sports_team |
| location | 17 | 778 | city, road, park |
| product | 14 | 43 | airplane, mobile_phone |
| art | 7 | 96 | film, written_work, music |
| event | 6 | 92 | election, natural_disaster |
| science | 10 | 46 | biology, chemistry |
| else | 8 | 89 | time, color, language |

Table 1: Concept type distribution of CR-LLM datasets. "#FGRC" denotes number of fine-grained reference concepts. And "#Num" denotes number of samples.

**Context Leakage Prevention** Context leakage occurs when sentences contain concepts of the masked entity. For example, the sentence "[entity] is an American singer" contains the reference concept "singer" for the masked entity. To address this issue, we first remove samples where the input included reference concepts. Additionally, since this is an open-ended question and the reference concept is not the only answer, we also perform manual filtering, in which we remove samples that contained any reasonable answer. For example, "[entity] is an American vocalist" with the reference concept "singer".

### 3.3 Dataset Analysis

In this section, we analyze the CR-LLM from four dimensions: dataset statistics, leakage prevention, solvability and diversity.

**Dataset Statistics** Based on FIGER, we create CR-LLM with 2,167 samples and covering 85 reference concepts. Each sample contains text with a masked entity and a reference concept. The average text length is 137.74 tokens, with 1.15 words per reference concept. Table 1 shows the concept type distribution. There are more samples about person, organization, and place, while the number of samples about other types is relatively smaller. To ensure the diversity of CR-LLM, each concept type contains at least 40 samples in CR-LLM. We also test the type distribution bias between CR-LLM and FIGER due to filtering operations in dataset construction. The results indicate that our filtering introduce only a minor distribution bias, with the largest deviation being just 4.32%, observed in the

| | Llama 2 | Vicuna | ChatGPT | Claude | GPT4 | Gemini |
|---|---|---|---|---|---|---|
| parameters | 7B | 7B | - | - | - | - |
| FIGER | 84.0 | 82.4 | 90.4 | 95.0 | 97.4 | 94.8 |
| MASKED | 12.0 | 20.4 | 28.8 | 26.0 | 44.8 | 34.0 |
| CR-LLM | 0.0 | 6.0 | 0.0 | 7.0 | 19.0 | 14.0 |

Table 2: Proportion of modeledge leakage in the FIGER, masked-FIGER and CR-LLM.

organization category.

**Leakage Prevention** To demonstrate the necessity and effectiveness of modeledge leakage prevention, we evaluate modeledge leakage in FIGER and CR-LLM. The results are shown in Table 2. In the original FIGER dataset, almost all models know the concepts of unmasked entities due to the fact that LLM retained a large amount of factual knowledge. Through our modeledge leakage prevention, effective control over modeledge leakage is achieved in CR-LLM. For small or medium-sized models, such as Llama2 (Touvron et al., 2023), Vicuna (Zheng et al., 2023) and Claude (Bai et al., 2022), data leakage was less than 7%. Larger models with more than 500B parameters, such as GPT-4 (OpenAI, 2023) and Gemini (Team et al., 2023), keep modeledge leakage within 20%. In addition, we examine modeledge leakage in masked-FIGER as ablation, which demonstrates that each step in our modeledge leakage prevention is necessary.

For context leakage, we manually analysis 500 randomly selected samples from FIGER. The results showed that about 5% of the samples contained reference concepts and about 32% contain other concepts of entity. And with our context leakage prevention, these context leakages are removed in CR-LLM.

We also investigate the adequacy of leakage prevention, which answers whether the filtering operation removes non-leakage cases. For modeledge leakage prevention, as both scenarios described in Section 3.2 inevitably lead to modeledge leakage, there is no risk of excessive filtering. For context leakage prevention, an examination of 100 randomly filtered samples revealed that 5% of the samples might be at risk of excessive filtering. However, considering that our dataset is used for testing, excessive filtering is acceptable compared to under-filtering.

**Solvability** Due to our filtering and modification of the FIGER dataset, there is a concern about the solvability of our data. To answer this concern, we asked 5 volunteers to perform concept reasoning on 200 random samples and count the number of unsolvables. The answers showed that 96% of the sample could be answered by the volunteers. And the remaining 4% could also be answered after knowledge of the relevant terminology was provided. Therefore, we claim that our CR-LLM dataset is solvable and usable.

| Role | Example | Proportion |
|---|---|---|
| ArgS | [entity] was obsessed with Barnabas Collins. | 11.0 |
| ArgO | Klehs lost a primary to [entity] in the 10th Senate District. | 35.0 |
| ArgM | Annappes is a village of France, on the [entity]. | 54.0 |

Table 3: Different roles of the masked entities in the sentence.

**Entity Diversity** Considering different roles of masked entities affect the difficulty of entity understanding, we also analysis the different roles of the masked entities in the text. We categorize the masked entities into three types according to their roles: ArgS (Agent and Theme), ArgO (other arguments) and ArgM (adjunct-like arguments) (Palmer et al., 2005; Roth and Lapata, 2016). Similarly, we randomly select 200 samples and use manual statistics on the proportions of the different types, and the results are shown in Table 3. Among them, ArgO and ArgM have more samples than ArgS, which may be caused by the fact that ArgS samples are more prone to modeledge leakage and context leakage. However, all three types of samples have sufficient numbers. Therefore, our dataset is complete and diverse in that it contains data of different roles and levels of difficulty.

### 3.4 Evaluation

Concept reasoning is an open-ended task, where there exists no particular ground truth. Therefore, we resort to GPT-4 for evaluation (Naismith et al., 2023; Xiao et al., 2024; Hackl et al., 2023).

Two types of metrics are used: *correctness* and *granularity*. For correctness, the reasonable answers related to the reference concept are considered as correct answers. We input the context, the unmasked entity, and the answer into GPT-4 to calculate accuracy, precision, recall, and macro-F1. For granularity, first, we test the number of words in reasoning results as length. Also, we

| Models | Accuracy | Precision | Recall | Macro-F1 | Length | Better-R | Worse-R | All-R | AQS |
|---|---|---|---|---|---|---|---|---|---|
| Llama-2-7b$_{fewshot}$ | 26.50 | 47.16 | 22.53 | 30.49 | 1.32 | 67.57 | 5.47 | 62.10 | 25.05 |
| Llama-2-7b$_{cot}$ | 37.00 | 59.93 | 40.21 | 48.13 | 1.57 | 57.09 | 14.18 | 42.90 | 31.75 |
| Vicuna-7b$_{fewshot}$ | 39.00 | 55.16 | 35.64 | 43.30 | 1.60 | 42.85 | 16.88 | 25.97 | 32.41 |
| Vicuna-7b$_{cot}$ | 51.91 | 69.01 | 47.82 | 56.49 | 1.41 | 35.80 | 32.09 | 3.70 | 35.25 |
| ChatGPT$_{fewshot}$ | 53.00 | 73.17 | 53.12 | 61.55 | 1.92 | 48.00 | 34.54 | 13.45 | 34.69 |
| ChatGPT$_{cot}$ | 75.00 | 87.54 | 75.12 | 80.86 | 1.10 | 42.10 | 23.68 | 18.42 | 57.24 |
| Gemini$_{fewshot}$ | 63.02 | 70.33 | 59.13 | 64.25 | 1.20 | 63.09 | 8.33 | 54.76 | 57.77 |
| Gemini$_{cot}$ | 77.44 | 86.69 | 78.71 | 82.51 | 1.29 | 48.42 | 15.78 | 32.63 | 65.22 |
| GPT4$_{fewshot}$ | 84.87 | 88.87 | 86.03 | 87.43 | 1.50 | 56.00 | 15.00 | 41.00 | 72.13 |
| GPT4$_{cot}$ | 85.29 | 90.65 | 86.58 | 88.57 | 1.64 | 56.00 | 10.00 | 46.00 | 76.76 |
| Model Effectiveness Score | 91 | - | - | - | - | - | - | 84 | - |

Table 4: Results of different models on CR-LLM dataset. "Better-R" ("worse-R") indicates the proportion of the predicted concept granularity superior (inferior) to the reference. "All-R"="Better-R"-"worse-R". "AQS" indicates the Accuracy-Quality Composite Score. And we computed the modeledge leakage fraction with the same accuracy as the non-modeledge leakage fraction to avoid scoring gaps due to modeledge from different models.

use the reference concept as a benchmark and use GPT-4 to compare the granularity of the reasoning results with the reference concept. Points are awarded for better concepts and deducted for worse concepts, called better_concepts_rate and worse_concepts_rate, providing a quality comparison between reasoning results and reference results. Finally, we also provide a composite score, which mixes accuracy with quality for evaluation, calculated as $accuracy * (1 - worse\_concepts\_rate)$.

Following the above evaluation methodology, we test the performance of five models of different sizes on CR-LLM and the results are presented in Table 4. And the results show that current models, especially those of common size and smaller scales, perform poorly in concept reasoning tasks. In addition, to validate the effectiveness of GPT-4, three volunteers provide scores for the scoring results, denoted as Model Effectiveness Score (MES). A score is assigned only when all three volunteers agree on the model evaluation. The high MES in Table 4 indicates that the above evaluation method is reasonable and effective.

# 4 Hybrid Method

As Table 4 shows that existing widely-used LLMs achieve limited performance on concept reasoning, we propose a hybrid method by combining inductive and deductive reasoning to stimulate the concept reasoning capability of LLMs. And the framework of the hybrid method is shown in Figure 2.

## 4.1 Inductive Reasoning

Inductive reasoning involves specific observation of patterns to draw a more general conclusion (Lawson, 2005). As suggested by many previous works (Wang et al., 2023; Rytting and Wingate, 2021; Olsson et al., 2022), concept reasoning based on inductive reasoning resorts to finding similar cases to the input sample and using their summaries to derive answers. Since the summarization is done by LLM, the key part here is to get a suitable similar cases. We use semantic similarity and literal similarity to identify similar cases. For semantic similarity, a direct method is to compute the cosine similarity between the input text and all other selectable cases, called global similarity. However, global similarity has limited performance due to variations in the roles of masked entities in sentences as shown in Table 3. Therefore, we additionally compute the cosine similarity between subclauses containing the masked entity, called local similarity. For literal similarity, different text similarity metrics can be adopted, such as BLEU (Papineni et al., 2002) and Jaccard similarity. Finally, we combine global and local semantic similarity to collect 5 most similar samples as candidates, then use local literal similarity to select 3 most similar samples from these candidate samples as final cases.

## 4.2 Deductive Reasoning

Deductive reasoning drives specific conclusions based on more general premises (Lawson, 2005). As suggested by many previous works (Ling et al., 2023a; Yan et al., 2023; Bostrom et al., 2022), con-

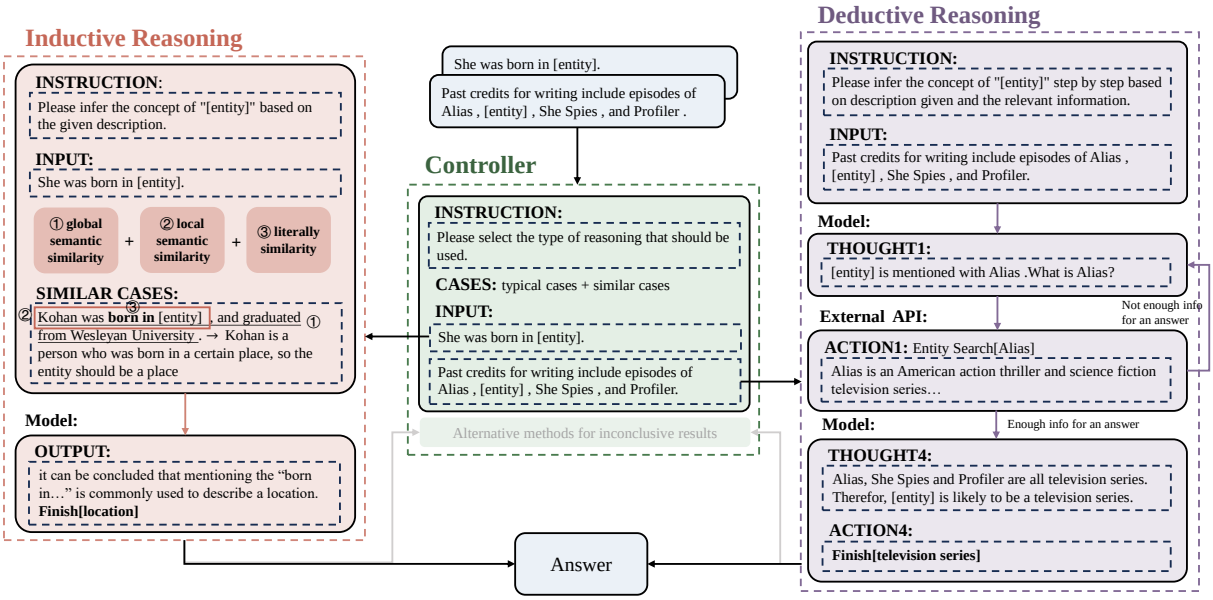


Figure 2: Workflow of the hybrid method, including inductive reasoning, deductive reasoning and a controller.

cept reasoning based on deductive reasoning resorts to iteratively using common sense in model and factual information, which can be thought of as premises, to draw conclusions. The process of deductive reasoning is illustrated on the right side of Figure 2. And we follow the ReAct format (Yao et al., 2022), in which we use the "thought" to infer the concept of the masked entity based on existing factual information and the "action" to retrieve factual information or outputs the final answer.

As factual information retrieval is the key to deductive reasoning, we introduce a retrieve-fallback-generate strategy to collect high-quality factual knowledge by combining knowledge base (Wikipedia) and LLM. The process starts by selecting a mention $m_j$ in the input and using an external knowledge base to retrieve knowledge about $m_j$. In cases the selected mention is related to multiple entities, we use a LLM to select the most appropriate entity. If the retrieval step encounters an error, such as no mention information in the external knowledge base, we resort to a LLM to directly generate the factual information.

### 4.3 Controller

In general, it is difficult to use a single reasoning method that accommodates diverse samples. Therefore, we propose an adaptive reasoning strategy that allows the LLM to automatically select the appropriate reasoning methods for different samples.

Specifically, the controller to determine the most appropriate reasoning method for the current input is a LLM using prompt and similar cases. For prompt, we specify how inductive and deductive reasoning work in concept reasoning. For the cases, we use two types of illustrative examples to assist the model's judgment. The first type consists of expert-selected exemplars for both types of reasoning, to help the model understand the characteristics of different reasoning methods. The second type selects the top-$K$ most similar cases based on cosine similarity metrics. Finally, to ensure the quality of output concept, if one reasoning method fails to provide an answer, we resort to the other reasoning method for output.

## 5 Experiments

In this section, we first conduct extensive experiments to verify the effectiveness of our hybrid method. Then, we provide a detailed analysis and case study of our method and current model in concept reasoning.

### 5.1 Experimental Setup

We compare our hybrid method with the following LLM reasoning methods: 1) Few-shot LLM. 2) In-Context Learning (cosine similarity). 3) Chain-of-Thought (Wei et al., 2022). 4) LLM with CoT+ICL. 5) Self Consistency (Wang et al., 2022): Reasoning 3 times and choose the answer with the highest number of occurrences. 6) Question Decomposition (Zhou et al., 2022): Decompose the question into sub-questions and provide answers. 7) Deductive Verification (Ling et al., 2023b): Perform

| | Methods | Accuracy | Precision | Recall | Macro-F1 | Length | Better-R | Worse-R | All-R | AQS |
|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | Few-shot LLM$_{llama}$ | 26.50 | 47.16 | 22.53 | 30.49 | **1.92** | 48.00 | 34.54 | 13.45 | 17.34 |
| | Few-shot LLM$_{chatgpt}$ | 53.00 | 73.17 | 53.12 | 61.55 | 1.32 | **67.57** | **5.47** | **62.10** | 50.10 |
| | In-Context Learning$_{llama}$ | 32.50 | 39.62 | 15.70 | 22.49 | 1.17 | 42.33 | 40.33 | 2.00 | 19.39 |
| | In-Context Learning$_{chatgpt}$ | 72.66 | 80.40 | 73.56 | 76.83 | 1.18 | 52.20 | 12.50 | 39.70 | 63.57 |
| | Chain-of-Thought$_{llama}$ | 37.00 | 59.93 | 40.21 | 48.13 | 1.10 | 42.10 | 23.68 | 18.42 | 28.23 |
| | Chain-of-Thought$_{chatgpt}$ | 75.00 | 87.54 | 75.12 | 80.86 | 1.57 | 57.09 | 14.18 | 42.90 | 64.36 |
| | LLM with ICL+CoT$_{llama}$ | 54.40 | 76.00 | 53.54 | 62.83 | 1.69 | 50.57 | 13.70 | 36.78 | 46.94 |
| | LLM with ICL+CoT$_{chatgpt}$ | 75.66 | 84.74 | 73.05 | 78.46 | 1.91 | 64.49 | 12.31 | 52.17 | 66.34 |
| | Self Consistency$_{llama}$ | 39.25 | 58.72 | 37.06 | 45.44 | 1.11 | 50.00 | **10.29** | 39.7 | 35.21 |
| | Self Consistency$_{chatgpt}$ | 75.33 | 85.66 | 74.08 | 79.45 | 1.58 | 59.35 | 14.02 | 45.32 | 64.76 |
| | Question Decomposition$_{llama}$ | 43.00 | 75.04 | 54.14 | 62.90 | 1.24 | 56.62 | 16.86 | 39.75 | 35.75 |
| | Question Decomposition$_{chatgpt}$ | 61.00 | 79.48 | 59.01 | 67.73 | 1.76 | 61.88 | 14.79 | 47.08 | 51.97 |
| | Deductive Verification$_{llama}$ | 44.60 | 71.34 | 50.24 | 58.96 | 1.19 | 48.91 | 16.30 | 32.60 | 37.33 |
| | Deductive Verification$_{chatgpt}$ | 79.00 | 85.71 | 75.86 | 80.49 | 1.63 | 56.41 | 15.54 | 40.87 | 66.72 |
| Ours | Inductive Reasoning$_{llama}$ | 55.60 | 75.75 | 55.00 | 63.73 | 1.66 | **66.25** | 13.75 | **52.50** | 47.95 |
| | Inductive Reasoning$_{chatgpt}$ | 79.33 | 85.49 | 75.29 | 80.07 | 1.93 | 68.55 | 12.01 | 56.53 | 69.80 |
| | Deductive Reasoning$_{llama}$ | 42.00 | 63.47 | 43.65 | 51.73 | 1.45 | 63.88 | 19.44 | 44.44 | 33.83 |
| | Deductive Reasoning$_{chatgpt}$ | 72.66 | 83.94 | 70.10 | 76.40 | **2.10** | 61.48 | 14.84 | 46.64 | 61.87 |
| | Hybrid Reasoning$_{llama}$ | **58.60** | **79.87** | **59.61** | **68.27** | 1.70 | 64.89 | 15.95 | 48.93 | **49.25** |
| | Hybrid Reasoning$_{chatgpt}$ | **82.00** | **89.22** | **81.50** | **85.18** | **2.10** | 67.22 | 12.70 | 54.51 | **71.58** |

Table 5: Results of different methods on CR-LLM dataset based on ChatGPT and Llama-2-7b. "Better-R" ("worse-R") indicates the proportion of the predicted concept granularity superior (inferior) to the reference. "All-R"="Better-R"-"worse-R". "AQS" indicates the Accuracy-Quality Composite Score.

self-verification for each reasoning. In addition, to validate the robustness of our method, we conduct experiments using both a closed-source large-parameter model, ChatGPT (OpenAI, 2022), and an open-source small-parameter model, Llama2-7B (Touvron et al., 2023).

## 5.2 Main Results

To validate the effectiveness of the hybrid method, we perform comparisons with baselines. The experimental results are shown in Table 5.

We conclude from the results: First, our method outperforms competitors in both accuracy and F1, showing that our method can accurately reason about the concept of the masked entity. Specifically, our hybrid method outperforms Deductive Verification by 3% and CoT+ICL by 6.5%. Second, the concepts generated by our method are longer and have more fine-grained data than most baselines. It should be noted that Few-shot LLM generates more fine-grained concepts than we do, which can be attributed to the fact that some of the results are guesses and LLMs are more likely to produce fine-grained concepts when making guesses. Third, within our methods, hybrid reasoning gives the best results compared to single reasoning, as different samples are better suited for different reasoning methods. Specifically, our method outperforms inductive reasoning in accuracy by 2.7%. And inductive reasoning shows better performance compared to deductive reasoning. One possible reason is that the logic of deductive reasoning is more complex and requires valid search targets to better motivate performance.

| | better-D | worse-D | All-D |
|---|---|---|---|
| Few-shot LLM | 20.28 | 31.88 | -11.59 |
| In-Context Learning | 16.01 | 38.28 | -22.26 |
| Chain-of-Thought | 12.40 | 28.19 | -15.78 |
| LLM with ICL+ CoT | 17.11 | 30.03 | -12.92 |
| Self Consistency | 16.66 | 42.85 | -26.19 |
| Question Decomposition | **28.57** | 47.61 | -19.04 |
| Deductive Verification | 16.66 | 43.33 | -26.66 |
| Inductive Reasoning | 3.70 | 33.33 | -29.62 |
| Deductive Reasoning | - | - | - |
| Hybrid Reasoning | 13.42 | **16.25** | **-2.80** |

Table 6: Differences in granularity between the results of other methods and deductive reasoning. "-D" means compared with deductive reasoning.

## 5.3 Detailed Analysis

In this subsection, we conduct experiments on other subtasks to further validate the effectiveness of our method. This includes the validity of deductive reasoning, sub-methods and hybrid strategies.

**Analysis of deductive reasoning** To show the advantage of deductive reasoning in concept quality, we compare the granularity differences between the

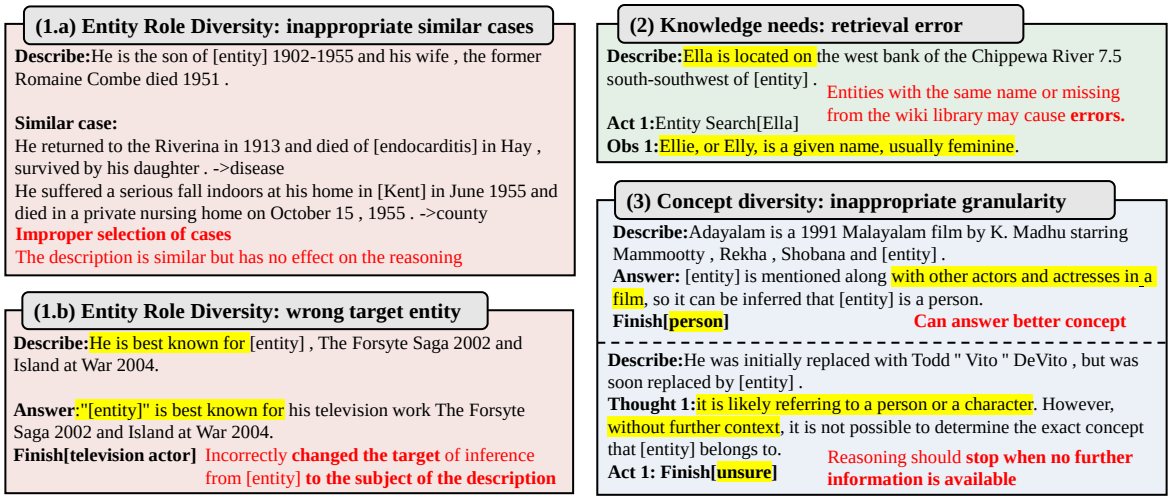


Figure 3: Some bad cases in concept reasoning and their reasons, which are divided into three parts: entity roles diversity, knowledge needs and concept diversity.

| | UB | Acc. | F1 | length | Qual. |
|---|---|---|---|---|---|
| CoT & Ind | 86.5 | 80.8 | 88.77 | 1.88 | -6.31 |
| CoT & Ded | 84.5 | 77.2 | 87.40 | 1.82 | -4.00 |
| SC & Ind | 86.2 | 80.0 | 81.96 | 1.87 | -6.25 |
| SC & Ded | 85.4 | 80.3 | 84.00 | 1.74 | -15.0 |
| QD & Ind | 83.4 | 78.8 | 84.80 | 1.99 | -8.79 |
| QD & Ded | 81.8 | 74.4 | 81.17 | 2.09 | -5.13 |
| DV & Ind | 87.0 | 79.3 | 82.26 | 1.77 | -6.12 |
| DV & Ded | 84.8 | 80.3 | 81.40 | 1.72 | -14.0 |
| Ind & Ded | **87.5** | **82.0** | **85.18** | **2.10** | - |

Table 7: Results of combining different sub-methods. UB denotes the upper bound and Qual denotes the granularity comparison with our hybrid reasoning.

| | accuracy | precision | recall | macro-F1 |
|---|---|---|---|---|
| Random Mix | 75.60 | 82.46 | 70.37 | 75.94 |
| Prompt-based | 78.00 | 87.37 | 77.73 | 82.27 |
| ICL-based | 79.33 | 87.34 | 77.29 | 82.01 |
| Ind2Ded | 80.80 | 87.33 | 78.71 | 82.80 |
| Ded2Ind | 76.33 | **89.61** | 75.78 | 82.12 |
| Bert Classifier | 77.00 | 88.11 | 74.92 | 80.98 |
| Ours | **82.00** | 89.22 | **81.50** | **85.18** |

Table 8: Results of different hybrid strategies. Prompt-based (ICL-based) means using prompts (similar cases) to classify. And Ind2Ded (Ded2Ind) means using inductive (deductive) reasoning and switches to deductive (inductive) reasoning when uncertain.

results of deductive reasoning and each baseline method, as shown in Table 6. From the table, we can see that all competitors have coarser granularity than deductive reasoning. This illustrates the importance of knowledge for the quality of concept reasoning. In addition, hybrid reasoning is closest to deductive reasoning in terms of granularity, but lags slightly behind due to the incorporation of inductive reasoning.

**Analysis of different sub-methods** To validate the effectiveness of combining the two sub-methods in hybrid reasoning, we systematically replace these two sub-methods with alternative baselines. The experimental results are shown in Table 7. The hybrid method based on induction and deduction shows the best performance in terms of both upper bounds and experimental results. This is due to the complementary nature of these two sub-methods. Although the other sub-methods maybe more accurate when executed individually, the lack of complementarity with inductive or deductive reasoning does not allow maximizing the benefits of hybrid.

**Analysis of different hybrid strategies** In this paper, we design a controller to select the appropriate reasoning method for different samples. To validate the effectiveness of this controller, we also design several classification methods as hybrid strategy for comparison. The results are shown in Table 8. And our controller outperforms almost all competitors on all metrics, highlighting the effectiveness of our controller.

## 5.4 Case Study

To gain a deeper understanding of the difficulty of concept reasoning for current models and methods, we present some cases as shown in Figure 3.

The first difficulty arises from the diversity of

entity roles. On the one hand, this makes it difficult to find suitable similar cases, which will affect the effectiveness of in-context learning, as in Figure 3 (1.a). On the other hand, the model may confuse the Agent with the masked entity, as in Figure 3 (1.b). Second, concept reasoning tasks require the model to have knowledge, either possessed by the model itself or returned by retrieval (as in our deductive reasoning), which may lead to errors because of incorrect knowledge. Finally, the diversity of concepts also adds difficulty to reasoning, which is shown in Figure 3 (3). Since too coarse granularity is not a good answer and pursuing too fine granularity may lead to errors, it is difficult for the model to decide whether to stop reasoning or continue and answer an optimal concept.

## 6 Conclusions

In this paper, we present the concept reasoning task for large language models for the first time. To evaluate the concept reasoning abilities of large language models, we construct a dataset, CR-LLM, which is built by two steps: modeledge leakage prevention and context leakage prevention. The dataset consists of 2,167 samples, covering different entity types, such as "person" and "place". Based on the dataset, we propose a hybrid reasoning method, consisting of inductive reasoning, deductive reasoning, and a controller. Experimental results show that LLMs perform sub-optimally in concept reasoning. In contrast, our proposed method significantly improves the concept reasoning abilities of LLMs, achieving a 7% increase in accuracy compared to CoT and demonstrating better conceptual granularity.

## Limitations

For our dataset, the reference concepts provided may not be optimal due to the use of previous entity typing datasets (e.g., FIGER). For example, if "US location" can be inferred from the text, the reference concept may be limited to "location" only. This could potentially reduce the quality of concept generation using the ICL method, given the presence of coarse-grained concepts in the examples. However, this does not affect our evaluation, since we use GPT-4 for evaluation.

There is room for improvement in our method. For example, self-consistency or verification enhancements could be considered. However, in this paper, we have highlighted our focus on combining complementary methods to be more effective than a single method. Therefore, we refrain from adding additional components to improve accuracy so as not to dilute our main focus.

## Ethical Considerations

This paper introduces a novel concept reasoning dataset constructed from the publicly available FIGER dataset. As the source dataset is openly accessible, there are no specific ethical considerations outlined in this paper.

## Acknowledgements

This work was supported by the Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902), the National Natural Science Foundation of China (No. 62306112) and Shanghai Sailing Program (No. 23YF1409400).

## References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer.

Haodong Bai, Frank Z Xing, Erik Cambria, and Win-Bin Huang. 2019. Business taxonomy construction using concept-level hierarchical clustering. *arXiv preprint arXiv:1906.09694*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Kaj Bostrom, Zayne Sprague, Swarat Chaudhuri, and Greg Durrett. 2022. Natural language deduction through search over statement compositions. *arXiv preprint arXiv:2201.06028*.

Jindong Chen, Yizhou Hu, Jingping Liu, Yanghua Xiao, and Haiyun Jiang. 2019. Deep short text classification with knowledge powered attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6252–6259.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. *arXiv preprint arXiv:1807.04905*.

Ning Ding, Yulin Chen, Xu Han, Guangwei Xu, Pengjun Xie, Hai-Tao Zheng, Zhiyuan Liu, Juanzi Li, and Hong-Gee Kim. 2021. Prompt-learning for fine-grained entity typing. *arXiv preprint arXiv:2108.10604*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. Is gpt-4 a reliable rater? evaluating consistency in gpt-4 text ratings. *arXiv preprint arXiv:2308.02575*.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *Preprint*, arXiv:2106.07139.

Anton E Lawson. 2005. What is the role of induction and deduction in reasoning and scientific inquiry? *Journal of Research in Science Teaching*, 42(6):716–740.

Chin Lee, Hongliang Dai, Yangqiu Song, and Xin Li. 2020. A chinese corpus for fine-grained entity typing. *arXiv preprint arXiv:2004.08825*.

Xiao Ling and Daniel Weld. 2012. Fine-grained entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 94–100.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023a. Deductive verification of chain-of-thought reasoning. *arXiv preprint arXiv:2306.03872*.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. 2023b. Deductive verification of chain-of-thought reasoning. *arXiv preprint arXiv:2306.03872*.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403.

Anh-Duong Nguyen, Kiem-Hieu Nguyen, and Van-Vi Ngo. 2019. Neural sequence labeling for vietnamese pos tagging and ner. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–5. IEEE.

Yuyang Nie, Yuanhe Tian, Yan Song, Xiang Ao, and Xiang Wan. 2020. Improving named entity recognition with attentive ensemble of syntactic information. *arXiv preprint arXiv:2010.15466*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

OpenAI. 2022. Chatgpt. https://openai.com/blog/chatgpt/.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

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.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022. Copen: Probing conceptual knowledge in pre-trained language models. *arXiv preprint arXiv:2211.04079*.

Ian Porada, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. Modeling event plausibility with consistent conceptual abstraction. *arXiv preprint arXiv:2104.10247*.

Sarah Masud Preum, Sile Shu, Homa Alemzadeh, and John A Stankovic. 2020. Emscontext: Ems protocol-driven concept extraction for cognitive assistance in emergency response. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13350–13355.

Jing Qiu, Yuhan Chai, Zhihong Tian, Xiaojiang Du, and Mohsen Guizani. 2019. Automatic concept extraction based on semantic graphs from big data in smart city. *IEEE Transactions on Computational Social Systems*, 7(1):225–233.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Michael Roth and Mirella Lapata. 2016. Neural semantic role labeling with dependency path embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1192–1202.

Christopher Rytting and David Wingate. 2021. Leveraging the inductive bias of large language models for abstract textual reasoning. *Advances in Neural Information Processing Systems*, 34:17111–17122.

Ritu Sharma, Dinesh Gopalani, and Yogesh Meena. 2017. Concept-based approach for research paper recommendation. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 687–692. Springer.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. 2023. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Ralph Weischedel and Ada Brunstein. 2005. Bbn pronoun coreference and entity type corpus. *Linguistic Data Consortium, Philadelphia*, 112.

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From automation to augmentation: Large language models elevating essay scoring landscape. *arXiv preprint arXiv:2401.06431*.

Shaotian Yan, Chen Shen, Junjie Liu, and Jieping Ye. 2023. Concise and organized perception facilitates large language models for deductive reasoning. *arXiv preprint arXiv:2310.03309*.

Xiyuan Yang, Xiaotao Gu, Sheng Lin, Siliang Tang, Yueting Zhuang, Fei Wu, Zhigang Chen, Guoping Hu, and Xiang Ren. 2019. Learning dynamic context augmentation for global entity linking. *arXiv preprint arXiv:1909.02117*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Mohamed Amir Yosef, Sandro Bauer, Johannes Hoffart, Marc Spaniol, and Gerhard Weikum. 2012. Hyena: Hierarchical type classification for entity names. In *Proceedings of COLING 2012: Posters*, pages 1361–1370.

Siyu Yuan, Deqing Yang, Jiaqing Liang, Zhixu Li, Jinxi Liu, Jingyue Huang, and Yanghua Xiao. 2022. Generative entity typing with curriculum learning. *arXiv preprint arXiv:2210.02914*.

Siyu Yuan, Deqing Yang, Jiaqing Liang, Jilun Sun, Jingyue Huang, Kaiyan Cao, Yanghua Xiao, and Rui Xie. 2021. Large-scale multi-granular concept extraction based on machine reading comprehension. In *International Semantic Web Conference*, pages 93–110. Springer.

Siyu Yuan, Deqing Yang, Jinxi Liu, Shuyu Tian, Jiaqing Liang, Yanghua Xiao, and Rui Xie. 2023. Causality-aware concept extraction based on knowledge-guided prompting. *arXiv preprint arXiv:2305.01876*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.