🕯️ CANDLE: Iterative Conceptualization and Instantiation Distillation from Large Language Models for Commonsense Reasoning

Weiqi Wang[1], Tianqing Fang[1], Chunyang Li[2], Haochen Shi[1], Wenxuan Ding[1], Baixuan Xu[1], Zhaowei Wang[1], Jiaxin Bai[1], Xin Liu[3], Jiayang Cheng[1], Chunkit Chan[1], Yangqiu Song[1]

[1]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
[2]Department of Computer Science and Technology, Tsinghua Univeirsity, Beijing, China
[3]Amazon.com Inc, Palo Alto, USA
{wwangbw, tfangaa, yqsong}@cse.ust.hk

Abstract

The sequential process of conceptualization and instantiation is essential to generalizable commonsense reasoning as it allows the application of existing knowledge to unfamiliar scenarios. However, existing works tend to undervalue the step of instantiation and heavily rely on pre-built concept taxonomies and human annotations to collect both types of knowledge, resulting in a lack of instantiated knowledge to complete reasoning, high cost, and limited scalability. To tackle these challenges, we introduce CANDLE (**C**onceptu**A**lization and I**N**stantiation **D**istillation from **L**arge Language Mod**E**ls), a distillation framework that iteratively performs contextualized conceptualization and instantiation over commonsense knowledge bases by instructing large language models to generate both types of knowledge with critic filtering. By applying CANDLE to ATOMIC (Sap et al., 2019a), we construct a comprehensive knowledge base comprising six million conceptualizations and instantiated commonsense knowledge triples. Both types of knowledge are firmly rooted in the original ATOMIC dataset, and intrinsic evaluations demonstrate their exceptional quality and diversity. Empirical results indicate that distilling CANDLE on student models provides benefits across three downstream tasks.

1 Introduction

Commonsense reasoning refers to the cognitive ability to make logical inferences and draw conclusions based on general knowledge and understanding of the world that is typically shared among individuals (Davis, 2014; Mueller, 2014). However, a longstanding challenge is generalizability, as commonsense reasoning often necessitates applying knowledge to novel situations beyond simple pattern recognition or memorizing all special cases (Mortimer, 1995; Banaji and Crowder, 1989). One promising approach to address this is the chain of conceptualization (Murphy, 2004) and instantiation (Anderson et al., 1976), which, akin to the process of conceptual induction and deduction in human reasoning (Tenenbaum et al., 2011), involves conceptualizing instances derived from known commonsense knowledge and subsequently instantiating these concepts in new situations to obtain the knowledge required for downstream reasoning. For example, in Figure 1, one can first conceptualize *enjoys exercising in the gym* as a *healthy lifestyle*, and then further instantiate it to *go on a balanced diet*. This process allows for the derivation of a novel event, *PersonX goes on a balanced diet*, which may entail new commonsense knowledge when connected with the original

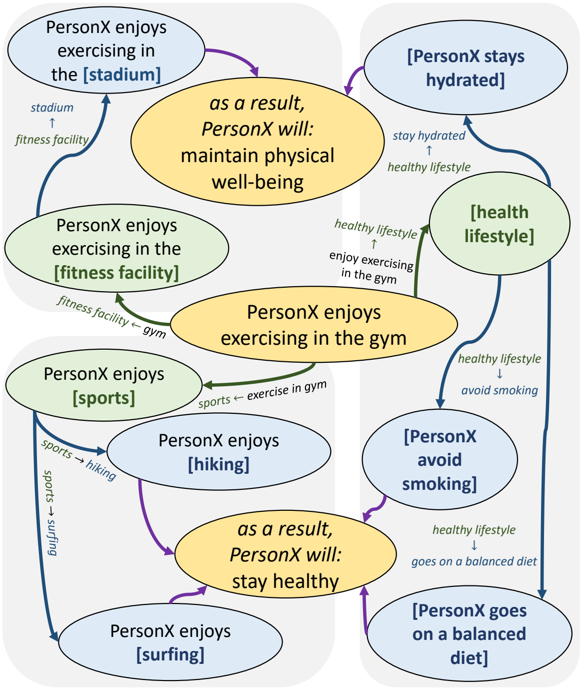
PersonX enjoys exercising in the [stadium]
as a result, PersonX will: maintain physical well-being
[PersonX stays hydrated]
stadium ↑ fitness facility
stay hydrated ↑ healthy lifestyle
PersonX enjoys exercising in the [fitness facility]
healthy lifestyle ↑ enjoy exercising in the gym
[health lifestyle]
fitness facility ← gym
PersonX enjoys exercising in the gym
PersonX enjoys [sports]
sports ← exercise in gym
healthy lifestyle ↓ avoid smoking
sports → hiking
PersonX enjoys [hiking]
[PersonX avoid smoking]
sports → surfing
as a result, PersonX will: stay healthy
healthy lifestyle ↓ goes on a balanced diet
PersonX enjoys [surfing]
[PersonX goes on a balanced diet]

Figure 1: Examples showing several chains of **conceptualization** and **instantiation** over the event *PersonX enjoys exercising in the gym*. **New inferential commonsense knowledge** can be induced when placing the **instantiation** back into the **original context**.

event's commonsense inferential tail. By possessing substantial knowledge to initiate the process of conceptualization and instantiation, one can extrapolate limited commonsense knowledge to a wide array of diverse scenarios.

Yet, replicating this fundamental ability on machines remains challenging due to the absence of both types of knowledge in widely used CommonSense Knowledge Bases (CSKBs; Sap et al., 2019a; Speer et al., 2017; Fang et al., 2021b,a). Various methods compensating the lack of conceptualization ability of language models have been proposed for entity-level (Durme et al., 2009; Song et al., 2011, 2015; Gong et al., 2016; He et al., 2020; Peng et al., 2022a) and event-level (Chen et al., 2020; He et al., 2024; Wang et al., 2023b) conceptualizations by matching against concept taxonomies like Probase (Wu et al., 2012) and WordNet (Miller, 1995). However, several limitations still persist.

Firstly, despite the importance of both conceptualization and instantiation, most existing works underestimate the importance of the second step while focusing solely on conceptualization and using the resulting abstract knowledge directly. Other studies that concentrate on instantiations either overlook the conceptualization step entirely or only retrieve instances from the original CSKB, failing to introduce novel entities and events. Secondly, most conceptualization methods heavily depend on matching instances with concepts in concept taxonomies, such as Probase and WordNet, which have a limited scope and lack contextual information. Consequently, the derived conceptualizations are constrained in scale by these taxonomies and are formulated without considering proper contextualization, necessitating further verification in the original context. Lastly, the chain of conceptualization and instantiation can easily bring more than two orders of magnitude of data on top of the original CSKB. However, current acquisition and verification methods for both steps heavily rely on human annotation, which can be extremely costly as the scale of the CSKB increases.

To address these gaps, we introduce CANDLE, a **C**onceptu**A**lization and I**N**stantiation **D**istillation framework from **L**arge Language Mod**E**ls (LLMs) to aid commonsense reasoning. Specifically, CANDLE marks the first to complete the chain of conceptualization and instantiation by instructing powerful LLMs to sequentially generate both types of knowledge based on concrete commonsense triples while carefully considering the original context throughout the process. We further alleviate the human annotation cost by employing two critic filtering models to eliminate low-quality generations. The instantiated knowledge, representing concrete commonsense knowledge again, can be fed back into CANDLE as input, iteratively augmenting the original CSKB significantly.

By applying CANDLE to ATOMIC (Sap et al., 2019a), we construct a large-scale knowledge base comprising 6.18 million conceptualizations and instantiations from two powerful LLMs, ChatGPT (OpenAI, 2022) and LLAMA2 (Touvron et al., 2023). We demonstrate the intrinsic efficacy of CANDLE through automatic and human evaluations, highlighting the ability to generate high-quality and diverse knowledge (Section 5.1). We further show the extrinsic benefits of CANDLE by leveraging the generated knowledge as complementary training data to distill student models that yield improvements across three downstream tasks, including CSKB conceptualization, generative commonsense inference, and zero-shot commonsense question answering (Section 5.2).

2 Related Works

2.1 Conceptualization and Instantiation

Conceptualization aims to abstract a set of entities or events into a general concept, thereby forming abstract commonsense knowledge within its original context (Murphy, 2004). Subsequently, instantiation grounds the derived concept into other instances and events to introduce new commonsense knowledge. Existing works primarily focused on entity-level conceptualization (Durme et al., 2009; Song et al., 2011, 2015; Liu et al., 2022; Peng et al., 2022a), with He et al. (2024) pioneering the construction of an event conceptualization benchmark by extracting concepts for social events from WordNet (Miller, 1995) synsets and Probase (Wu et al., 2012). Wang et al. (2023b,a) further proposed a semi-supervised framework for conceptualizing CSKBs and demonstrated that abstract knowledge can enhance commonsense inference modeling and question answering. Wang et al. (2023d) constructed an abstraction benchmark based on eventualities from ASER (Zhang et al., 2022). Regarding instantiation, Allaway et al. (2023) introduced a controllable generative framework to identify valid instantiations for abstract knowledge automatically. However, none of the

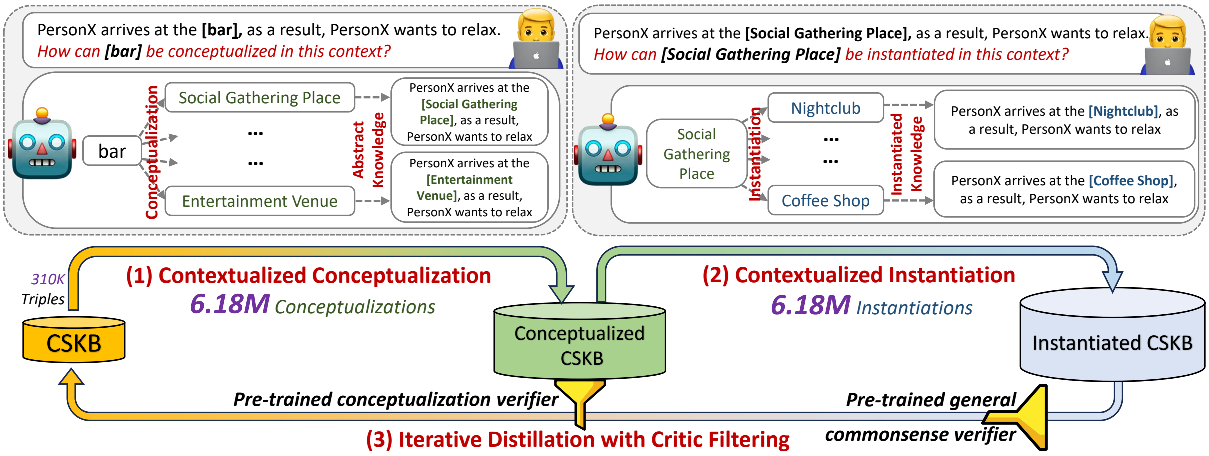
PersonX arrives at the [bar], as a result, PersonX wants to relax.
How can [bar] be conceptualized in this context?
bar
Conceptualization
Social Gathering Place
...
...
Entertainment Venue
Abstract Knowledge
PersonX arrives at the [Social Gathering Place], as a result, PersonX wants to relax
PersonX arrives at the [Entertainment Venue], as a result, PersonX wants to relax
PersonX arrives at the [Social Gathering Place], as a result, PersonX wants to relax.
How can [Social Gathering Place] be instantiated in this context?
Social Gathering Place
Instantiation
Nightclub
...
...
Coffee Shop
Instantiated Knowledge
PersonX arrives at the [Nightclub], as a result, PersonX wants to relax
PersonX arrives at the [Coffee Shop], as a result, PersonX wants to relax
310K Triples
CSKB
(1) Contextualized Conceptualization
6.18M Conceptualizations
Conceptualized CSKB
(2) Contextualized Instantiation
6.18M Instantiations
Instantiated CSKB
Pre-trained conceptualization verifier
Pre-trained general commonsense verifier
(3) Iterative Distillation with Critic Filtering

Figure 2: Overview of our CANDLE framework. A running example with *PersonX arrives at the bar, as a result, PersonX wants to relax* is shown in the figure, where *bar* is first conceptualized and then instantiated by LLMs. The instantiations can be integrated back into the original CSKB and become input for the framework again.

existing studies have fully completed the chain of conceptualization and instantiation, with each focusing on only one aspect. Human annotation is also frequently applied for data collection and verification, which is both expensive and limited in scalability. Additionally, the downstream benefits of instantiated commonsense knowledge have not been thoroughly explored, leaving a significant gap in improving commonsense reasoning models.

2.2 Commonsense Knowledge Distillation

Recent breakthroughs in LLMs (OpenAI, 2022, 2023) have led to numerous efforts in distilling commonsense knowledge into datasets for training performant student models. West et al. (2022); Sclar et al. (2022); Bhagavatula et al. (2023); West et al. (2023) followed the pipeline of symbolic knowledge distillation, which uses human-crafted prompts to extract specific types of knowledge from LLMs for training downstream models. He et al. (2022) proposed to transfer distilled knowledge from a ranker to a retriever, resulting in a more robust commonsense generator. Chae et al. (2023) and Kim et al. (2023) focused on distilling conversational responses from LLMs to enhance dialogue agents with commonsense knowledge and high-quality rationales. In this paper, we share similar aspirations and propose a chain of distillation framework that sequentially obtains abstract and instantiated knowledge from powerful LLMs. Empirical results show that our framework offers more substantial downstream benefits than traditional symbolic knowledge distillation methods.

3 Definitions and Datasets

We follow the definitions proposed by He et al. (2024) and Wang et al. (2023b) to formulate conceptualization and instantiation. Denote the triples in the original CSKB as $D_o = \{(h_o, r, t) | h_o \in H_o, r \in R, t \in T\}$, where H_o, R, and T are the set of heads, relations, and tails in the original CSKB. The objective of conceptualization is to form a conceptualized head event, denoted as h_a, from the original head h_o. This is achieved by linking a component $i \subseteq h_o$ to a concept c, forming h_a by replacing i with c. Consequently, abstract knowledge is formed by combining the conceptualized head event with the original relation and tail, represented by (h_a, r, t). In the next step, the goal of instantiation is to associate the concept $c \subseteq h_a$ with a new instance i'. This process enables the formation of new commonsense knowledge in the format of $(h_{i'}, r, t)$, where $h_{i'}$ is obtained by replacing $c \subseteq h_a$ with i'. In this paper, we use ATOMIC (Sap et al., 2019a) as the original CSKB D_o, which contains 310K (h_o, r, t) triples after dropping those with wildcards and 18,839 unique h_o head events. AbstractATOMIC (He et al., 2024) is used as the source of instances i for every head event h_o.

4 CANDLE

This section introduces our CANDLE framework, illustrated in Figure 2. Our framework can be outlined in three steps: (1) Instruct ChatGPT to generate contextualized conceptualizations based on the triples in the original CSKB. (2) Instruct LLAMA2 to instantiate the conceptualizations obtained in

Step 1. (3) Apply critic-filtering to the generations in both steps and close the loop by reintroducing the instantiations back to the CSKB.

4.1 Contextualized Conceptualization

Previous methods for collecting conceptualizations rely on heuristically matching instances against concepts from WordNet and Probase. However, they suffer from limited concept coverage, resulting in a lack of knowledge diversity after instantiation, and require additional verification to ensure that concept c fits into the original context (h_o, r, t). To address both issues, we propose to utilize ChatGPT as a loose teacher to collect conceptualizations in a one-step inference manner. To verify the feasibility of such choice instead of other open-source LLMs, we carry out a pilot study, in which we randomly select 1000 events and asked ChatGPT and LLAMA2-7B to generate their conceptualizations. The results of our expert evaluation show that ChatGPT has 98% plausible generations, while LLAMA2 only achieves 81%. Therefore, we choose ChatGPT as our core conceptualizer due to its exceptional performances. Following Brown et al. (2020) and West et al. (2022), we use a few-shot prompt to instruct ChatGPT:

```
<TASK-PROMPT>
<EX$_1$-INP><EX$_1$-OUT>
...
<EX$_{N-1}$-INP><EX$_{N-1}$-OUT>
<EX$_N$-INP>
```

where **<TASK-PROMPT>** is a task instruction that explains how to conceptualize an event and **<EX$_1$-INP><EX$_1$-OUT>** are human authored examples of conceptualizations for events sampled from ATOMIC. For each example, (h_o, r, t, i) are included in the input, and c is the output. Finally, we provide the N$_{\text{th}}$ input as **<EX$_N$-INP>** and ask ChatGPT to generate the corresponding conceptualization as **<EX$_N$-OUT>**. This ensures that ChatGPT not only learns the relationship between instances i and their conceptualizations c but also performs such abstraction in a contextualized manner, ensuring the plausibility of the generated conceptualization c within the original context (h_o, r, t). In this paper, we set $N = 6$ and obtain $N_c = 20$ conceptualizations for every event h_o.

4.2 Contextualized Instantiation

After conceptualizing all events, we proceed to instantiate them by instructing an open-source LLM to reduce the cost as the scale of instantiation is $N_c = 20$ times larger than that of conceptualization. Similarly, we carry out a round of pilot study to demonstrate the feasibility of employing LLAMA2. We ask both ChatGPT and LLAMA2 to instantiate 1,000 ChatGPT-generated conceptualizations, and find that both models are able to produce approximately 95% plausible instantiations with critic filtering. Considering the significant cost of using ChatGPT to generate 6.18 million conceptualizations, we decide to use LLAMA2-13B as our core instantiater (Case studies are shown in Appendix G). We employ a similar prompt as described in Section 4.1, with the modification of replacing **<TASK-PROMPT>** with the explanation of instantiating a conceptualized event and changing **<EX$_1$-INP><EX$_1$-OUT>** to human-authored examples of instantiations for abstract commonsense knowledge triples. (h_a, r, t, c) are included in the input and i' is the expected output. By learning from these examples, LLAMA2 is expected to generate the corresponding instantiation i' (**<EX$_N$-OUT>**) based on the given abstract knowledge triple (h_a, r, t, c) (**<EX$_N$-INP>**). We set $N = 11$ and produce only one instantiation for each conceptualized event h_a due to the significant amount of conceptualizations obtained in the previous step. Appendix A provides more details regarding the distillation process.

4.3 Iterating with Critic Filtering

Following West et al. (2022), we use critic filtering models to eliminate low-quality generations from LLMs. Specifically, we utilize a DeBERTa-v3-large conceptualization discriminator, provided by Wang et al. (2023b), and VERA-T5-xxl, provided by Liu et al. (2023), to evaluate the quality of the generated conceptualizations and instantiations, respectively. We set an empirical threshold value t to serve as the cutoff point for discarding generations with scores below t. In Section 5.1, we present evaluations conducted to determine the optimal value for t. For all downstream applications, we set $t = 0.9$. Post-filtering, the instantiated triples $(h_{i'}, r, t)$ can be reintroduced as the input for conceptualizations again as they continue to represent concrete commonsense knowledge. This iterative process of conceptualization and instantiation forms a loop, which enables continuously augmenting a CSKB. In this paper, we execute the loop only once, but multiple iterations hold

Corpus	Conceptualization		Instantiation	
	Size (Unq.)/K	Accept	Size (Unq.)/K	Accept
AbsATM	503.5 (31.22)	-	None	-
EXEM	0.650 (0.650)	-	25.12 (25.12)	-
CANDLE	6,181 (853.5)	82.6%	6,181 (676.7)	77.9%
(critic$_{0.5}$)	4,002 (498.4)	88.1%	4,176 (512.7)	84.4%
(critic$_{0.7}$)	3,272 (382.2)	93.5%	3,098 (455.9)	89.1%
(critic$_{0.9}$)	2,137 (219.4)	97.2%	2,208 (382.1)	94.5%

Table 1: Statistics and expert acceptance rates of CANDLE in comparison to AbstractATOMIC (AbsATM; He et al., 2024) and Exemplar (EXEM; Allaway et al., 2023). Unq stands for unique.

the promise of significantly enhancing the CSKB's knowledge coverage.

5 Evaluations and Analysis

In this section, we evaluate CANDLE from both intrinsic and extrinsic perspectives. Intrinsically, we demonstrate the high quality and diversity of conceptualizations and instantiations generated by CANDLE (Section 5.1). Extrinsically, we explore the benefits by applying the distilled knowledge to downstream tasks (Section 5.2).

5.1 Distillation Evaluations

Statistics and Quality. We present CANDLE distillation statistics based on ATOMIC in Table 1, showing its superiority in scale and concept coverage compared to other benchmarks. Even with a strict critic filtering threshold ($t = 0.9$), CANDLE maintains its leading position, having the highest count of total and unique knowledge for both types. To assess the quality of the distilled knowledge, we recruit four expert annotators to conduct human evaluations on the plausibility of the generated conceptualizations and instantiations. They are asked to annotate the plausibility of 3,000 randomly sampled abstract commonsense triples (h_a, r, t) and 3,000 instantiated triples $(h_{i'}, r, t)$ from the distilled knowledge set. Accepted triples are those deemed plausible by all annotators. We then analyze accepted triple ratios for different levels of critic filtering, as shown in Table 1. Our findings show that LLMs have impressive conceptualization and instantiation abilities, with initial plausibility rates of 82.6% and 77.9% for both types of knowledge, respectively. Critic filtering improves plausibility by up to 14.6% and 16.6%, demonstrating the effectiveness of our measures in maintaining high-quality distilled knowledge. For more annotation details, refer to Appendix E.

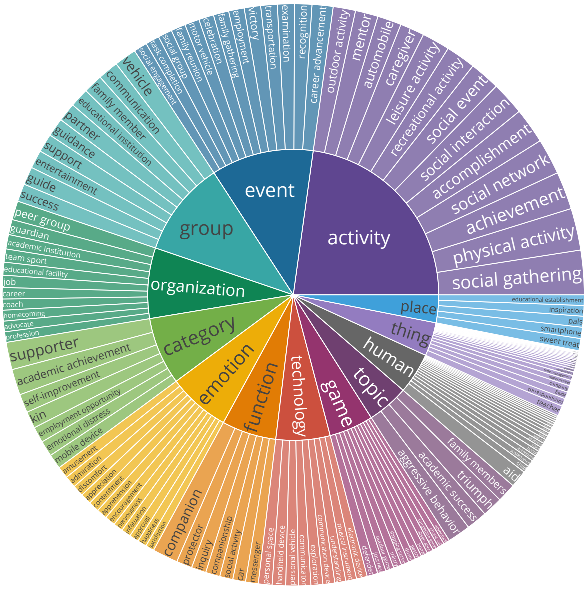

Figure 3: Hypernyms distribution of the top 10,000 popular conceptualizations distilled from CANDLE.

Conceptualization Diversity. The process of abstracting an event into highly diverse conceptualizations plays a crucial role in CANDLE. It is of significant importance because the greater the diversity of conceptualizations, the broader the knowledge coverage becomes upon instantiation. This, in turn, enhances the overall knowledge coverage within the distillation process. To examine the diversity of the top 10,000 popular distilled conceptualizations, we obtain their hypernyms by matching against Probase (Wu et al., 2012) and present a visualization in Figure 3. It reveals that our distilled conceptualizations possess a high level of diversity across various categories, forming a comprehensive and intricate knowledge base. Novelty of our distilled conceptualizations and instantiations, compared to other available resources, are in Appendix A.

5.2 Downstream Applications

In this section, we explore the downstream applications of CANDLE. By applying CANDLE to ATOMIC, the distilled conceptualizations and instantiations form a large-scale expansion of the original CSKB, which contains high-quality abstract and concrete commonsense knowledge. Leveraging both types of knowledge as supplementary training data, we enhance various downstream commonsense reasoning models. Specifically, we utilize distilled conceptualizations in the CSKB conceptualization task (Wang et al., 2023b), while instantiations are used in generative commonsense inference (COMET; Bosselut et al., 2019) and zero-shot commonsense QA tasks (Ma et al.,

Model Type	Backbone Model / Method	Event Conceptualization		Triple Conceptualization	
		Validation	Testing	Validation	Testing
Pre-trained Langauge Models	RoBERTa-large *340M*	77.28	77.99	81.77	82.69
	DeBERTa-v3-large *435M*	78.02	78.27	82.18	82.96
	GPT2-XL *1.5B*	53.71	56.10	47.65	47.21
	PseudoReasoner (RoBERTa-large)	78.33	78.91	79.69	80.27
	PseudoReasoner (DeBERTa-v3-large)	79.03	79.21	79.89	80.07
	CAT (RoBERTa-large) *340M*	78.51	78.53	82.27	83.02
	CAT (DeBERTa-v3-large) *435M*	$\underline{79.55}$	$\underline{79.39}$	$\underline{82.88}$	$\underline{83.52}$
Large Language Models	ChatGPT (`openai/gpt-3.5-turbo`)	69.29	68.65	68.54	68.12
	+ Five-shot Exemplars	69.42	70.40	70.27	72.08
	+ Chain-of-thought	74.82	72.32	71.48	72.85
	LLAMA2 *7B*	46.29	43.90	40.81	41.25
	+ Five-shot Exemplars	47.92	44.89	74.67	76.80
	LLAMA2 *13B*	48.17	48.59	48.31	48.55
	+ Five-shot Exemplars	49.29	49.90	$\underline{80.67}$	82.08
	Mistral-v0.1 *7B*	46.29	43.90	58.09	58.07
	+ Five-shot Exemplars	51.00	50.06	65.09	69.80
	LLAMA2 (LoRA Fine-tuned) *7B*	$\underline{75.80}$	76.27	79.89	$\underline{82.15}$
	Mistral-v0.1 (LoRA Fine-tuned) *7B*	75.71	$\underline{76.76}$	79.59	80.35
	VERA-T5 *5B*	70.76	70.29	72.60	76.85
	VERA-T5 (Fine-tuned) *5B*	75.69	76.21	80.13	81.25
CANDLE Distilled *(Ours)*	RoBERTa-large *340M*	$80.69_{\uparrow 2.18}$	$80.99_{\uparrow 2.46}$	$83.11_{\uparrow 0.84}$	$84.50_{\uparrow 1.48}$
	DeBERTa-v3-large *435M*	$\underline{\mathbf{80.97}}_{\uparrow 1.42}$	$\underline{\mathbf{81.14}}_{\uparrow 1.75}$	$\underline{\mathbf{83.64}}_{\uparrow 0.76}$	$\underline{\mathbf{84.64}}_{\uparrow 1.12}$
	LLAMA2 (LoRA Fine-tuned) *7B*	$77.48_{\uparrow 1.68}$	$78.27_{\uparrow 2.00}$	$81.68_{\uparrow 1.79}$	$83.40_{\uparrow 1.25}$
	Mistral-v0.1 (LoRA Fine-tuned) *7B*	$77.77_{\uparrow 2.06}$	$78.29_{\uparrow 1.53}$	$81.95_{\uparrow 2.36}$	$82.54_{\uparrow 2.19}$
	VERA-T5 (Fine-tuned) *5B*	$77.54_{\uparrow 1.85}$	$78.03_{\uparrow 1.82}$	$82.79_{\uparrow 2.66}$	$83.61_{\uparrow 2.36}$

Table 2: Performances (Accuracy%) on CSKB conceptualization tasks. The best performances within each model type are underlined, and the best among all models are **bold-faced**.

2021). Due to space constraints, please refer to Appendix B, C, D for task setups, dataset descriptions, and implementation details, respectively.

5.2.1 CSKB Conceptualization

Task Setup. The CSKB conceptualization task evaluates a model's ability to conceptualize a CSKB through two binary classification subtasks, which are crucial for performing CSKB conceptualization inference upon concept taxonomies (He et al., 2024). The first subtask, event conceptualization, aims to determine whether h_o can be correctly conceptualized using h_a, where h_a is derived by replacing an instance $i \subset h_o$ with its linked concept c. The second subtask, triple conceptualization, aims to assess the plausibility of a conceptualized triple (h_a, r, t) that represents abstract commonsense knowledge. Accuracy is used as the evaluation metric. Following Wang et al. (2023b), we use the AbstractATOMIC dataset provided by He et al. (2024) as the evaluation benchmark.

To obtain our distilled models for these tasks, we first synthesize negative samples from CANDLE distilled conceptualizations. For event conceptualizations, a random concept from another head event without common words is selected as the negative candidate, while for triple conceptualization, a tail of another head event without common words under the same relation is selected. We then fine-tune language models on a balanced mixture of CANDLE distillation and synthesized negative samples to train two models, each serving as a pre-trained general discriminator in their respective task domain. These two models are subsequently fine-tuned on the training sets of AbstractATOMIC to fit into the benchmark, and their performances on the validation and test sets are reported.

Baselines. We evaluate our distilled models by comparing them against several baselines. These include supervised fine-tuned language models like RoBERTa-Large (Liu et al., 2019), DeBERTa-V3-Large (He et al., 2023), GPT-2 (Radford et al., 2019), LLAMA2 (Touvron et al., 2023), Mistral (Jiang et al., 2023), and VERA (Liu et al., 2023), as well as semi-supervised methods such as PsuedoReasoner (Fang et al., 2022) and CAT (Wang et al., 2023b). Due to computational power limitations, we utilize LoRA (Hu et al., 2022) for fine-tuning LLMs. As additional baselines, we also consider prompting LLMs, including LLAMA2, Mistral, and ChatGPT. We explore both direct zero-shot prompting and alternative methods, such as with five-shot exemplars (Wei et al., 2023) and chain-of-thought reasoning (Wei et al., 2022).

Training Data	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L	CIDEr	BERTScore	Human
Backbone: GPT2-XL (Radford et al., 2019) ***1.5B***									
Zero-shot	4.350	1.598	0.732	0.293	5.702	5.030	0.792	37.11	14.50
ATOMIC	45.72	29.18	21.12	16.15	29.97	49.69	64.61	76.09	70.50
ATOMIC$^{20}_{20}$	42.15	25.77	17.82	13.14	29.82	47.61	63.70	70.39	76.50
ATOMIC-10X	45.38	29.20	21.09	16.15	30.09	49.86	65.02	75.89	77.50
AbstractATOMIC	45.30	29.08	21.00	16.06	29.98	48.61	63.98	75.56	71.50
CANDLE Distilled	**50.71**	**33.85**	**25.55**	**20.43**	**32.45**	**51.91**	**69.68**	**76.86**	78.50
Backbone: ChatGPT (OpenAI, 2022) (`openai/gpt-3.5-turbo`)									
Zero-shot	11.82	4.258	1.891	0.926	13.87	13.73	4.350	49.28	78.50
Five-shot	26.32	12.50	7.160	4.415	18.60	24.65	8.313	58.69	**81.00**
Chain-of-thought	9.906	3.568	1.556	0.736	11.85	11.02	2.905	46.17	64.00
Backbone: LLAMA2 (Touvron et al., 2023) ***7B***									
Zero-shot	18.26	7.453	3.594	1.945	15.90	20.28	8.872	48.23	48.50
Five-shot	31.22	16.87	9.767	5.989	19.74	27.67	17.83	58.41	65.50
ATOMIC	42.04	23.01	14.10	9.125	27.80	42.90	53.17	71.52	68.50
ATOMIC$^{20}_{20}$	41.07	22.46	13.62	8.619	27.74	42.42	53.28	71.77	74.00
ATOMIC-10X	42.07	23.08	14.14	9.198	28.14	42.75	53.69	71.93	76.50
AbstractATOMIC	42.78	23.64	14.58	9.471	27.74	42.55	53.12	71.51	71.00
CANDLE Distilled	43.86	24.40	15.12	10.00	28.36	43.86	54.25	72.94	79.50

Table 3: Performances (%) of the commonsense inference modeling task (COMET) on the full test set of ATOMIC$^{20}_{20}$. The best ones within each backbone are underlined, and the best among all is **bold-faced**.

Results and Analysis. Table 2 shows the results. CAT trained with DeBERTa-v3-large outperforms all other baselines for both tasks. Among LLMs, LLAMA and Mistral perform well after fine-tuning, but they struggle in prompting scenarios. However, pre-training on CANDLE's distilled conceptualizations consistently improves results for both tasks. For example, Mistral shows a significant improvement of 1.54% and 2.19% on two tasks compared to directly fine-tuning on AbstractATOMIC. Additionally, the distilled DeBERTa-v3-large surpasses all baseline models and achieves state-of-the-art performance. This can be attributed to the distilled conceptualizations obtained from CANDLE, which grant the model a more comprehensive understanding of conceptualizations and subsequently enhance its discriminatory capabilities.

5.2.2 Generative Commonsense Inference

Task Setup. The task of generative commonsense inference modeling (COMET; Bosselut et al., 2019) asks the model to generate commonsense tails t based on given head h_o and relation r inputs. Following Hwang et al. (2021), we use the full test set of ATOMIC$^{20}_{20}$ as our evaluation benchmark. We use several automatic metrics for evaluation, including BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Lavie and Agarwal, 2007), CIDEr (Vedantam et al., 2015), and BERTScore (Zhang et al., 2020). Meanwhile, four expert annotators are recruited to conduct expert evaluations of the generations. They are asked to annotate the plausibility of 200 randomly selected commonsense triple generations under each setting, and the resulting plausibility rates are reported.

Similar to training distilled models in previous tasks, we first pre-train GPT2 and LLAMA2-7B on critic-filtered CANDLE instantiations, where each $(h_{i'}, r, t)$ triple is concatenated into a sentence via natural language templates. Subsequently, we fine-tune these models on the training split of ATOMIC$^{20}_{20}$ to fit them into the benchmark. Finally, we report their performances on the test set.

Baselines. For baselines, we separately train GPT2 and LLAMA2-7B on the training sets of ATOMIC, ATOMIC$^{20}_{20}$, ATOMIC10X (West et al., 2022), and AbstractATOMIC. These models are then fine-tuned on the training split of ATOMIC$^{20}_{20}$ and evaluated on its test set. We also include their zero-shot prompting performances, with LLAMA2 being evaluated with five-shot exemplars. ChatGPT's performances under zero-shot, five-shot, and chain-of-thought settings are also reported.

Results and Analysis. Table 3 shows the results. Among the baselines, models pre-trained on ATOMIC-10X achieve the highest expert acceptance rate, surpassing those trained on AbstractATOMIC. This may be because ATOMIC-10X covers a wider range of commonsense relations consistent with ATOMIC$^{20}_{20}$. However, CANDLE distilled models achieve the highest scores compared to baselines with the same backbone model. For example, the CANDLE distilled LLAMA-7B model improves BERTScore by 1.01% and expert-plausibility by 3.00% compared to the best baseline. It also outperforms ChatGPT in all automatic met-

Model/Method	CSKB	a-NLI	CSQA	PIQA	SIQA	WG	Avg.
Pre-trained Language Models							
RoBERTa-L (Liu et al., 2019)	-	65.5	45.0	67.6	47.3	57.5	56.6
DeBERTa-v3-L (He et al., 2023)	-	59.9	25.4	44.8	47.8	50.3	45.6
Self-talk (Shwartz et al., 2020)	-	-	32.4	70.2	46.2	54.7	-
SMLM (Banerjee and Baral, 2020)	*	65.3	38.8	-	48.5	-	-
COMET-DynGen (Bosselut et al., 2021)	ATOMIC	-	-	-	50.1	-	-
MICO (Su et al., 2022)	ATOMIC	-	44.2	-	56.0	-	-
STL-Adapter (Kim et al., 2022)	ATOMIC	71.3	66.5	71.1	64.4	60.3	66.7
DeBERTa-v3-L (MR) (Ma et al., 2021)	ATM10X	75.1	71.6	79.0	59.7	71.7	71.4
DeBERTa-v3-L (MR) (Ma et al., 2021)	ATOMIC	76.0	67.0	78.0	62.1	76.0	71.8
CAR-DeBERTa-v3-L (Wang et al., 2023a)	ATOMIC	78.9	67.2	78.6	63.8	78.1	73.3
CAR-DeBERTa-v3-L (Wang et al., 2023a)	AbsATM	$\underline{79.6}$	69.3	78.6	64.0	$\underline{78.2}$	$\underline{73.9}$
DeBERTa-v3-L (CANDLE Distilled)	CANDLE	**81.2**$_{\uparrow 1.6}$	69.9$_{\uparrow 0.6}$	80.3$_{\uparrow 1.7}$	65.9$_{\uparrow 1.9}$	**78.3**$_{\uparrow 0.1}$	**74.9**$_{\uparrow 1.0}$
Large Language Models							
GPT-3.5 (`text-davinci-003`)	-	61.8	68.9	67.8	68.0	60.7	65.4
ChatGPT (`gpt-3.5-turbo`)	-	69.3	74.5	75.1	69.5	62.8	70.2
+ Chain-of-thought	-	70.5	$\underline{75.5}$	79.2	**70.7**	63.6	71.9
+ Self-consistent chain-of-thought	-	73.2	**75.7**	$\underline{81.7}$	$\underline{69.7}$	64.1	72.9
GPT-4 (`gpt-4`)	-	75.0	43.0	73.0	57.0	77.0	65.0
LLAMA2 (7B; Touvron et al., 2023)	-	57.5	57.8	78.8	48.3	69.2	62.3
LLAMA2 (13B; Touvron et al., 2023)	-	55.9	67.3	80.2	50.3	72.8	65.3
Mistral-v0.1 (7B; Jiang et al., 2023)	-	51.0	59.6	**83.0**	42.9	75.3	62.4
VERA-T5-xxl (Liu et al., 2023)	ATOMIC	71.2	61.7	76.4	57.7	67.5	66.9
VERA-T5-xxl (Liu et al., 2023)	ATM10X	70.3	59.5	75.1	58.2	67.2	66.1
VERA-T5-xxl (Liu et al., 2023)	AbsATM	73.2	63.0	77.2	58.1	68.1	68.0
VERA-T5-xxl (CANDLE Distilled)	CANDLE	73.8$_{\uparrow 0.6}$	64.7$_{\uparrow 1.7}$	77.6$_{\uparrow 0.4}$	59.4$_{\uparrow 1.2}$	71.3$_{\uparrow 3.2}$	69.4$_{\uparrow 1.4}$

Table 4: Zero-shot evaluation results (Accuracy%) on five commonsense question answering benchmarks. The best results are **bold-faced**, and the second-best ones are underlined. ATM10X stands for ATOMIC-10X (West et al., 2022) and AbsATM stands for AbstractATOMIC (He et al., 2024).

rics while maintaining a high plausibility rate of around 80%. This emphasizes the advantages of using CANDLE distilled instantiations for COMET training over traditional symbolic knowledge distillation methods or conceptualization augmentation. Interestingly, we also observe that LLAMA2 has a tendency to generate long and contextually rich commonsense knowledge. On the other hand, GPT2, when fine-tuned on ATOMIC-like data, may generate shorter and more concise knowledge, which aligns with the format and length of knowledge in ATOMIC2020, thus achieving better results in automatic evaluations. However, human annotators tend to consider long and contextually rich commonsense statements, generated by LLAMA2, as more plausible.

5.2.3 Zero-shot Commonsense QA

Task Setup. The task of zero-shot commonsense QA involves selecting the most plausible option for commonsense questions without any supervision signals from benchmark data. We follow the most effective pipeline by Ma et al. (2021), which fine-tune language models on QA pairs synthesized from knowledge in CSKBs. The head h_o and relation r of a (h_o, r, t) triple are transformed into a question using natural language prompts, with the tail t serving as the correct answer option. Distractors or negative examples are generated by randomly sampling tails from triples that do not share common keywords with the head. In addition to directly synthesizing from knowledge triples in ATOMIC, we augment ATOMIC by sampling triples from ATOMIC-10X, AbstractATOMIC, and CANDLE instantiations. The number of sampled triples is the same as in the original ATOMIC dataset. We then synthesize them into QA pairs to train different baseline models and CANDLE distilled models. For our distilled models, we utilize QA pairs sourced from CANDLE-instantiation augmented ATOMIC to train a DeBERTa-v3-large model using the marginal ranking loss and a T5-xxl model (Raffel et al., 2020) following the training regime of VERA. We evaluate the performance of all models on the validation split of Abductive NLI (aNLI; Bhagavatula et al., 2020), CommonsenseQA (CSQA; Talmor et al., 2019), PhysicalIQA (PIQA; Bisk et al., 2020), SocialIQA (SIQA; Sap et al., 2019b), and WinoGrande (WG; Sakaguchi et al., 2021). Accuracy is used as the evaluation metric.

Baselines. First, we report performances of vanilla RoBERTa-Large, DeBERTa-v3-Large, Self-talk (Shwartz et al., 2020), COMET-

Critic	Conceptualization	Instantiation
0.0	92.3%	85.5%
0.5	94.6%	91.2%
0.7	95.9%	93.3%
0.9	98.3%	96.7%

Table 5: Annotation results of distillations obtained from the second round of executing CANDLE.

DynaGen (Bosselut et al., 2021), SMLM (Banerjee and Baral, 2020), MICO (Su et al., 2022), MR (Ma et al., 2021), STL-Adapter (Kim et al., 2022), and the previous state-of-the-art method, CAR (Wang et al., 2023a). For MR and CAR, DeBERTa-v3-Large is used as the backbone, and their performances on ATOMIC-10X and AbstractATOMIC are also reported. For LLMs, we report the performances of prompting GPT3.5 (Brown et al., 2020), ChatGPT, GPT4 (OpenAI, 2023), LLAMA2, and Mistral in a zero-shot manner. For ChatGPT, its performances with chain-of-thought (Wei et al., 2022) and self-consistency chain-of-thought (Wang et al., 2023c) prompting are also reported. We also train several VERA-T5-xxl baselines on different sets of QA pairs as LLM baselines.

Results and Analysis. Table 4 shows the results, demonstrating that CANDLE distilled models generalize better than the baselines across several commonsense QA benchmarks. For instance, VERA demonstrates an average improvement of 1.4% compared to the best baseline. This can be attributed to the inclusion of new entities and events in CANDLE instantiations that are absent in other CSKBs, where CANDLE instantiations can aid in answering commonsense questions that require knowledge of these new instances. Furthermore, the distilled DeBERTa-v3-large model outperforms all baselines, including methods utilizing LLMs. This also indicates that augmenting with CANDLE distilled instantiations provides a more significant advantage compared to using symbolically distilled or abstract knowledge as training data.

5.3 Analysis

5.3.1 Feasibility of Iterating CANDLE

We first demonstrate the feasibility of iterating the CANDLE framework with more than one round. To do so, we randomly sample 10,000 distilled instantiations from LLAMA2 as the input for the CANDLE framework and execute the framework again, resulting in 200,000 conceptualizations and 200,000 instantiations. Subsequently, we randomly select 300 from each set and annotate them accordingly. The results are shown in Table 5. We observe that iterating the framework produces slightly better results than the first loop. This improvement may be attributed to the fact that the knowledge generated in the initial loop is more easily understood by LLMs compared to the human-annotated data in ATOMIC. Moreover, 58% of the conceptualizations and 44% of the instantiations are novel compared to the first loop. Based on these findings, we believe that our iterative framework is effective, and the iteration process enhances the augmentation of a CSKB through multiple iterations.

Task	CSKB Concept.	COMET	CSQA
Overlap Ratio	10.1%	8.7%	5.3%
Avg. Similarity	0.39	0.38	0.31

Table 6: Knowledge overlap ratio and average similarity between distilled knowledge and evaluation data.

5.3.2 Source of Empirical Gains

Since LLAMA2 has been pre-trained on some evaluation benchmarks, it remains questionable whether the empirical gains in downstream tasks are due to knowledge overlap between distillations from LLAMA2 and the evaluation benchmarks. To this extent, we further demonstrate that CANDLE distilled models perform better due to improved generalizability rather than relying on data overlap with the evaluation data. We use Sentence-BERT (Reimers and Gurevych, 2019) to measure the textual similarity between the distilled knowledge and the evaluation data for each task. We then calculate the ratio of data that exhibits semantic overlap with a similarity score exceeding 0.5 and also report the average similarity. The results are shown in Table 6. Based on the results, we observe that the distilled knowledge has minimal overlap with the evaluation set. This indicates that the empirical gain primarily stems from our distilled knowledge, which improves the generalizability of the models, rather than relying on knowledge overlap with the evaluation sets.

6 Conclusion

This paper introduces CANDLE, a distillation framework that realizes the chain of conceptualization and instantiation over CSKBs. We demonstrate the efficacy of CANDLE through comprehensive evaluations of the distilled knowledge and its positive impact on downstream tasks. Our research sheds light on distilling LLMs to enable more robust and generalizable commonsense reasoning.

Limitations

The major limitation of CANDLE lies in the significant cost of distilling LLMs to obtain substantial knowledge. While the instantiation step of CANDLE utilizes the open foundation model LLAMA2, the conceptualization is still performed by ChatGPT due to the unsatisfactory performance of other open-source LLMs and the high quality of ChatGPT's generation. Consequently, a considerable amount of funding is required to distill conceptualizations for CANDLE to function effectively. Future works should address this issue by exploring advanced prompting methods or employing stronger open-source LLMs as the foundation for distilling conceptualizations.

Furthermore, it should be noted that CANDLE has only been validated on ATOMIC. However, CANDLE is not limited to any specific format of commonsense knowledge, allowing it to operate on any CSKB (Shen et al., 2023). Future research can address this by extending the evaluation of CANDLE to other CSKBs and conducting follow-up experiments to explore their benefits on more downstream tasks.

Another interesting direction to investigate is utilizing the chain of conceptualization and instantiation as a foundation for enhancing weak-to-strong generalization (Burns et al., 2023). By conceptualizing and instantiating weak supervision data, we can generate more robust and generalized training signals, which ultimately strengthens the learning process. This can also be effectively incorporated into the training process of self-rewarding language models (Yuan et al., 2024).

Ethics Statement

To avoid generating harmful or unethical content from LLMs like ChatGPT and LLAMA2, we recruit four expert annotators, who are graduate and undergraduate students specializing in machine commonsense in natural language processing, to verify the ethics and potential harm of the generated content. A thorough assessment of a random sample has been conducted, and no significant harm has been identified. All training and evaluation datasets used are publicly available and shared under open-access licenses solely for research purposes, aligning with their intended usage. These datasets have been carefully anonymized and desensitized to protect data privacy and confidentiality. The expert annotators involved in this study are fully aware of the annotation protocol and the intended use of their annotations. Their participation in this research is voluntary, and they have agreed to contribute without receiving any compensation. Thus, the authors believe that this paper does not raise any ethical concerns.

Acknowledgements

We thank the anonymous reviewers and the area chair for their constructive comments. The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20), and the GRF (16211520 and 16205322) from RGC of Hong Kong. We also thank the support from the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21, RMGS23CR05, RMGS23EG08).

References

Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen R. McKeown, Doug Downey, and Yejin Choi. 2023. Penguins don't fly: Reasoning about generics through instantiations and exceptions. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2610–2627. Association for Computational Linguistics.

Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, and Yuxiong He. 2022. Deepspeed- inference: Enabling efficient inference of transformer models at unprecedented scale. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis, Dallas, TX, USA, November 13-18, 2022*, pages 46:1–46:15. IEEE.

Richard C Anderson, James W Pichert, Ernest T Goetz, Diane L Schallert, Kathleen V Stevens, and Stanley R Trollip. 1976. Instantiation of general terms. *Journal of Verbal Learning and Verbal Behavior*, 15(6):667–679.

Jiaxin Bai, Xin Liu, Weiqi Wang, Chen Luo, and Yangqiu Song. 2023. Complex query answering on eventuality knowledge graph with implicit logical constraints. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.*

Mahzarin R Banaji and Robert G Crowder. 1989. The bankruptcy of everyday memory. *American Psychologist*, 44(9):1185.

Pratyay Banerjee and Chitta Baral. 2020. Self-supervised knowledge triplet learning for zero-shot question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 151–162. Association for Computational Linguistics.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Lianhui Qin, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. 2023. I2D2: inductive knowledge distillation with neurologic and self-imitation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9614–9630. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Antoine Bosselut, Ronan Le Bras, and Yejin Choi. 2021. Dynamic neuro-symbolic knowledge graph construction for zero-shot commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 4923–4931. AAAI Press.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *CoRR*, abs/2312.09390.

Hyungjoo Chae, Yongho Song, Kai Tzu-iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5606–5632. Association for Computational Linguistics.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024a. Exploring the potential of chatgpt on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 684–721. Association for Computational Linguistics.

Chunkit Chan, Cheng Jiayang, Yauwai Yim, Zheye Deng, Wei Fan, Haoran Li, Xin Liu, Hongming Zhang, Weiqi Wang, and Yangqiu Song. 2024b. Negotiationtom: A benchmark for stress-testing machine theory of mind on negotiation surrounding. *CoRR*, abs/2404.13627.

Muhao Chen, Hongming Zhang, Haoyu Wang, and Dan Roth. 2020. What are you trying to do? semantic typing of event processes. In *Proceedings of the 24th Conference on Computational Natural Language Learning, CoNLL 2020, Online, November 19-20, 2020*, pages 531–542. Association for Computational Linguistics.

Jiayang Cheng, Lin Qiu, Tsz Ho Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. 2023. Storyanalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11518–11537. Association for Computational Linguistics.

Ernest Davis. 2014. *Representations of commonsense knowledge*. Morgan Kaufmann.

Zheye Deng, Chunkit Chan, Weiqi Wang, Yuxi Sun, Wei Fan, Tianshi Zheng, Yauwai Yim, and Yangqiu Song. 2024. Text-tuple-table: Towards information integration in text-to-table generation via global tuple extraction. *CoRR*, abs/2404.14215.

Zheye Deng, Weiqi Wang, Zhaowei Wang, Xin Liu, and Yangqiu Song. 2023. Gold: A global and local-aware denoising framework for commonsense knowledge graph noise detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3591–3608. Association for Computational Linguistics.

Wenxuan Ding, Shangbin Feng, Yuhan Liu, Zhaoxuan Tan, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Knowledge crosswords: Geometric reasoning over structured knowledge with large language models. *CoRR*, abs/2310.01290.

Benjamin Van Durme, Phillip Michalak, and Lenhart K. Schubert. 2009. Deriving generalized knowledge from corpora using wordnet abstraction. In *EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, Athens, Greece, March 30 - April 3, 2009*, pages 808–816. The Association for Computer Linguistics.

Tianqing Fang, Zeming Chen, Yangqiu Song, and Antoine Bosselut. 2024. Complex reasoning over logical queries on commonsense knowledge graphs. *CoRR*, abs/2403.07398.

Tianqing Fang, Quyet V. Do, Hongming Zhang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2022. Pseudoreasoner: Leveraging pseudo labels for commonsense knowledge base population. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3379–3394. Association for Computational Linguistics.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. Benchmarking commonsense knowledge base population with an effective evaluation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8949–8964. Association for Computational Linguistics.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yu Gong, Kaiqi Zhao, and Kenny Qili Zhu. 2016. Representing verbs as argument concepts. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2615–2621. AAAI Press.

Xin Guan, Biwei Cao, Qingqing Gao, Zheng Yin, Bo Liu, and Jiuxin Cao. 2023. Multi-hop commonsense knowledge injection framework for zero-shot commonsense question answering. *CoRR*, abs/2305.05936.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2024. Acquiring and modeling abstract commonsense knowledge via conceptualization. *Artificial Intelligence*, page 104149.

Mutian He, Yangqiu Song, Kun Xu, and Dong Yu. 2020. On the role of conceptualization in commonsense knowledge graph construction. *CoRR*, abs/2003.03239.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Xingwei He, Yeyun Gong, A-Long Jin, Weizhen Qi, Hang Zhang, Jian Jiao, Bartuer Zhou, Biao Cheng, Sm Yiu, and Nan Duan. 2022. Metric-guided distillation: Distilling knowledge from the metric to ranker and retriever for generative commonsense reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 839–852. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12930–12949. Association for Computational Linguistics.

Yu Jin Kim, Beong-woo Kwak, Youngwook Kim, Reinald Kim Amplayo, Seung-won Hwang, and Jinyoung Yeo. 2022. Modularized transfer learning with multiple knowledge graphs for zero-shot commonsense reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2244–2257. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, pages 228–231. Association for 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.

Jiacheng Liu, Wenya Wang, Dianzhuo Wang, Noah Smith, Yejin Choi, and Hannaneh Hajishirzi. 2023. Vera: A general-purpose plausibility estimation model for commonsense statements. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1264–1287, Singapore. Association for Computational Linguistics.

Jingping Liu, Tao Chen, Chao Wang, Jiaqing Liang, Lihan Chen, Yanghua Xiao, Yunwen Chen, and Ke Jin. 2022. Vocsk: Verb-oriented commonsense knowledge mining with taxonomy-guided induction. *Artif. Intell.*, 310:103744.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Feihong Lu, Weiqi Wang, Yangyifei Luo, Ziqin Zhu, Qingyun Sun, Baixuan Xu, Haochen Shi, Shiqi Gao, Qian Li, Yangqiu Song, and Jianxin Li. 2024. MIKO: multimodal intention knowledge distillation from large language models for social-media commonsense discovery. *CoRR*, abs/2402.18169.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. 2021. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13507–13515. AAAI Press.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Eduardo F Mortimer. 1995. Conceptual change or conceptual profile change? *Science & Education*, 4:267–285.

Erik T Mueller. 2014. *Commonsense reasoning: an event calculus based approach*. Morgan Kaufmann.

Gregory Murphy. 2004. *The big book of concepts*. MIT press.

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

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. 2022a. COPEN: probing conceptual knowledge in pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5015–5035. Association for Computational Linguistics.

Xiangyu Peng, Siyan Li, Sarah Wiegreffe, and Mark Riedl. 2022b. Inferring the reader: Guiding automated story generation with commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7008–7029, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social iqa: Commonsense reasoning about social interactions. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 4462–4472. Association for Computational Linguistics.

Melanie Sclar, Peter West, Sachin Kumar, Yulia Tsvetkov, and Yejin Choi. 2022. Referee: Reference-free sentence summarization with sharper controllability through symbolic knowledge distillation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9649–9668. Association for Computational Linguistics.

Xiangqing Shen, Siwei Wu, and Rui Xia. 2023. Dense-atomic: Towards densely-connected ATOMIC with high knowledge coverage and massive multi-hop paths. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13292–13305. Association for Computational Linguistics.

Haochen Shi, Weiqi Wang, Tianqing Fang, Baixuan Xu, Wenxuan Ding, Xin Liu, and Yangqiu Song. 2023. QADYNAMICS: training dynamics-driven synthetic QA diagnostic for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15329–15341. Association for Computational Linguistics.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4615–4629. Association for Computational Linguistics.

Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. 2011. Short text conceptualization using a probabilistic knowledgebase. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 2330–2336. IJCAI/AAAI.

Yangqiu Song, Shusen Wang, and Haixun Wang. 2015. Open domain short text conceptualization: A generative + descriptive modeling approach. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 3820–3826. AAAI Press.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Ying Su, Zihao Wang, Tianqing Fang, Hongming Zhang, Yangqiu Song, and Tong Zhang. 2022. MICO: A multi-alternative contrastive learning framework for commonsense knowledge representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1339–1351. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA,* *June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319, Dublin, Ireland. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: Conceptualization-augmented reasoner for zero-shot commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13520–13545, Singapore. Association for Computational Linguistics.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13111–13140. Association for Computational Linguistics.

Weiqi Wang and Yangqiu Song. 2024. MARS: Benchmarking the metaphysical reasoning abilities of language models with a multi-task evaluation dataset. *CoRR*, abs/2406.02106.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023c. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zhaowei Wang, Haochen Shi, Weiqi Wang, Tianqing Fang, Hongming Zhang, Sehyun Choi, Xin Liu, and Yangqiu Song. 2023d. Abspyramid: Benchmarking the abstraction ability of language models with a unified entailment graph. *CoRR*, abs/2311.09174.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *CoRR*, abs/2303.03846.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4602–4625. Association for Computational Linguistics.

Peter West, Ronan Le Bras, Taylor Sorensen, Bill Yuchen Lin, Liwei Jiang, Ximing Lu, Khyathi Chandu, Jack Hessel, Ashutosh Baheti, Chandra Bhagavatula, and Yejin Choi. 2023. Novacomet: Open commonsense foundation models with symbolic knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1127–1149. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November*

16-20, 2020, pages 38–45. Association for Computational Linguistics.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 481–492. ACM.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1173–1191. Association for Computational Linguistics.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *CoRR*, abs/2401.10020.

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2022. ASER: towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *Artif. Intell.*, 309:103740.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM.

Appendices

A Distillation Details

This section provides additional details about the CANDLE distillation process not covered in the main body text. First, we present the prompts used to instruct ChatGPT to perform contextualized conceptualizations and LLAMA2 to perform contextualized instantiation. For prompting ChatGPT to distill conceptualizations, we use a few-shot prompt as shown below:

```
Following the given examples, you are
required to conceptualize the instance
(enclosed by []) in the last given event
into abstract concepts. The concept
should still fit into the instance's
original sentence. Make sure that
the generated abstract concepts are
general and not simply hypernyms of the
instance.
...
Event <i>: PersonX enjoys drinking in
the [bar], as a result, PersonX feels
relaxed. [bar] can be conceptualized as
Social Gathering Place
...
Event <N>: PersonX likes [painting on
the beach], as a result, PersonX will go
to the beach. [painting on the beach]
can be conceptualized as
```

Similarly, for prompting LLAMA2-13B to distill instantiations based on previously generated conceptualizations, we use a few-shot prompt as shown below:

```
Following the given examples, you are
required to instantiate the concept
(enclosed by []) in the last given
event into entities or events. If the
event only contains the concept, then
instantiate it to an event starting
with a subject PersonX or PersonY. If
the event contains other words, then
instantiate it to an entity. The
instance should still fit into the
original sentence. Make sure that the
generated instance is specific.
...
Event <i>: PersonX enjoys drinking in
the [Social Gathering Place], as a
result, PersonX feels relaxed. [Social
Gathering Place] can be instantiated as
beer festival
...
Event <N>: PersonX likes [exercise], as
a result, PersonX will go to the stadium.
[exercise] can be conceptualized as
```

These prompts are consistent with our descriptions in Section 4.1 and Section 4.2, where the task description is first presented, followed by human-authored examples, and finally, the event we want to conceptualize or instantiate. We also leverage several tricks in the prompt, such as numbering the examples, generating concepts instead of hypernyms, and keeping the generated responses concise.

	Abs.ATM	CANDLE
#Unq. event	15,388	15,359
#Unq. instance	21,493	21,442
#Unq. conceptualization	31,227	**853,499**
#Tot. conceptualization	503,588	**6,181,391**
#Unq. instantiation	-	**676,737**
#Tot. instantiation	-	**6,181,391**
Avg. #concept/event	32.73	**173.33**
Avg. #Unq. concept/event	28.33	**167.76**
Avg. #concept/instance	23.43	**124.16**
Avg. #Unq. concept/instance	17.27	**100.88**

Table 7: Statistics of conceptualizations and instantiations in AbstractATOMIC (Abs.ATM; He et al., 2024) and CANDLE. Tot. stands for total, Unq. stands for unique, and Avg. stands for average.

Relation	ATOMIC	Abs.ATM	CANDLE
xEffect	78,832	938,330	964,765
oEffect	28,351	333,845	346,363
xWant	101,249	1,170,835	1,322,810
oWant	43,079	484,570	551,391
xReact	62,969	510,476	480,259
oReact	26,570	224,706	208,538
xNeed	74,272	900,429	894,338
xAttr	110,791	838,191	810,958
xIntent	45,490	519,813	601,969
Total	572,053	5,921,195	6,181,391

Table 8: Statistics of abstract commonsense knowledge triples by relations in ATOMIC, AbstractATOMIC (Abs.ATM; He et al., 2024), and CANDLE.

Finally, we parse the generations via manually defined rules and compile them into a dataset.

Additionally, we introduce some generation settings when prompting LLMs. For ChatGPT, we access it through the official OpenAI APIs[1]. The code of the accessed version is `gpt-3.5-turbo-0613`. We set the temperature to 1.0 and the maximum length for generated tokens to 200. To conceptualize all events in ATOMIC into 20 conceptualizations each, the time required for the distillation process is approximately ten days and the financial budget is around 1500 USD.

For LLAMA2, we access it via the Huggingface Library (Wolf et al., 2020). The code of the accessed model is `meta-llama/Llama-2-13b-chat-hf`[2]. When prompting, we use the Top-k sampling decoding strategy and set $k = 10$. We set the maximum length of generated tokens to 200. The models are hosted on sixteen NVIDIA-V100 GPUs, and the time required to distill the entire dataset is approximately one month.

After collecting 20 conceptualizations for every head event in ATOMIC and further instantiating them to new entities and events, we construct an expanded knowledge base of ATOMIC. We also include more statistics, as shown in Table 7 and Table 8. For instantiations, they share the same relational distribution as abstract commonsense triples since we only instantiate them once. These statistics indicate that, compared to AbstractATOMIC, which is the only available conceptualization benchmark based on ATOMIC, CANDLE contains more abstract commonsense triples and many more unique conceptualizations. According to our results, it can also be expected that the abstract knowledge distilled from CANDLE is of better quality than AbstractATOMIC, which human annotations or any filtering have not verified.

For critic filtering, we use the state-of-the-art conceptualization discriminator developed by Wang et al. (2023b). This discriminator is utilized to assess the plausibility of CANDLE distilled conceptualizations. It considers the original event, the instance being conceptualized, and the target concept as its inputs and generates a score ranging from 0 to 1 to represent plausibility. For instantiation, we use the pre-trained VERA model released by Liu et al. (2023). We convert the instantiated commonsense knowledge triple into a declarative statement and request an estimation of its plausibility from VERA. This estimation is provided as a score ranging from 0 to 1. The output scores from both models serve as the critical values assigned to each CANDLE distillation. These critical values are then subjected to further filtering based on various thresholds.

Additionally, following Wang et al. (2023d), we calculate the percentage of unique abstract concepts using BLEU soft uniqueness (Zhu et al., 2018; West et al., 2022). We define a concept, denoted as x, as unique if $BLEU_1(C, x) < 0.5$, where C represents all concepts that share the same head event and identified instance with x in AbstractATOMIC. Here, 0.5 serves as an empirical threshold. Our distillation process yields 92.3% unique conceptualizations, indicating a significantly higher diversity than previous datasets.

Similarly, we evaluate the uniqueness of the newly introduced head events resulting from our chain of conceptualization and instantiation. To

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

[2] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf

determine uniqueness, we define an instantiated head event, referred to as $h_{i'}$, as unique if $BLEU_1(h_o, h_{i'}) < 0.5$, where h_o represents the original head event in ATOMIC. The threshold of 0.5 is an empirical threshold. Our empirical results demonstrate that 78.6% of the instantiated events are unique compared to ATOMIC, highlighting the effectiveness of CANDLE in enhancing the semantic coverage of the CSKB.

Data	Type	Train	Dev	Test
D^l	#event	107,384	12,117	11,503
	#triple	65,386	8,403	7,408
D^u	#event	304,983	36,023	31,578
	#triple	4,851,272	499,523	570,400

Table 9: Statistics of labeled data D^l and unlabeled data D^u in AbstractATOMIC.

B Task Setups

B.1 CSKB Conceptualization

We follow the task definition of He et al. (2024) and Wang et al. (2023b) to formulate the CSKB conceptualization task. Specifically, conceptualizing an event-centric CSKB to derive abstract commonsense knowledge comprises two steps (He et al., 2024): event conceptualization and triple conceptualization, which correspond to two subtasks studied in this paper. Denote the triples in the original CSKB as $D_o = \{(h_o, r, t) | h_o \in H_o, r \in R, t \in T\}$, where H_o, R, and T are the set of heads, relations, and tails in the original CSKB. The first step only operates on head events without considering the context in r and t. The goal of event conceptualization is to produce a conceptualized head event h_a from the original head h_o to represent an abstraction of h_o. In the second step, the task is to verify whether the conceptualized head h_a still makes sense in the context of r and t, as r and t will further restrict the level of abstractness in h_a. Plausible (h_a, r, t) triples will be considered as valid abstract commonsense knowledge. By enhancing the performance of discriminative models on these tasks, they can function as more precise critic filters and automate the conceptualization process of a CSKB when linked to concept taxonomies.

B.2 Generative Commonsense Inference

The task of generative commonsense inference was studied by both Bosselut et al. (2019) and Hwang et al. (2021). It requires a generative model to complete the tail t of a commonsense assertion based on a given pair of head h and commonsense relation r. In this paper, we follow Hwang et al. (2021) and use ATOMIC$^{20}_{20}$ as the evaluation benchmark, in which the full testing set is used for model evaluation. The task of COMET is important in the domain of commonsense as it serves as a fundamental component for numerous high-level applications that necessitate commonsense reasoning, such as zero-shot commonsense question answering with self-talk (Shwartz et al., 2020) and dynamic graph construction (Bosselut et al., 2021), narrative reasoning (Peng et al., 2022b), and dialogue generation (Tu et al., 2022). Improving COMET can potentially benefit other domains that require commonsense understanding.

B.3 Zero-shot Commonsense QA

The task of zero-shot commonsense QA evaluates a model's reasoning generalizability on unseen QA entries without any supervision signals from the corresponding annotated training data. Several methods have been proposed to tackle this task, including those by Shwartz et al. (2020); Bosselut et al. (2021); Kim et al. (2022); Shi et al. (2023). The most effective pipeline, as suggested by Ma et al. (2021), injects commonsense knowledge into language models via fine-tuning on QA pairs synthesized from knowledge in CSKBs. During the fine-tuning process, the head h_o and relation r of a (h_o, r, t) triple from a CSKB are transformed into a question using natural language prompts, with the tail t serving as the correct answer option. Distractors or negative examples are generated by randomly sampling tails from triples that do not share common keywords with the head. This fine-tuning procedure enhances the model's knowledge not only for QA benchmarks constructed from CSKBs but also improves its ability to answer unseen commonsense questions in a more generalized manner. In this paper, we follow the task definition, model training, and model evaluation pipeline by Ma et al. (2021) to study the impact of distilling student models from CANDLE instantiations. For baselines, we compare models trained on QA pairs synthesized from ATOMIC, ATOMIC-10X, and AbstractATOMIC. For ATOMIC-10X, 0.9 is used as the critic filtering threshold.

C Dataset Descriptions

This section covers additional details and statistics of datasets and benchmarks used in downstream task evaluations.

C.1 CSKB Conceptualization

In CSKB Conceptualization tasks, we use the AbstractATOMIC (He et al., 2024) dataset as the evaluation benchmark. It is a benchmark dataset built upon ATOMIC and consists of event conceptualization data and abstract knowledge triples. The event conceptualizations are based on head events in ATOMIC, identified through syntactic parsing and matching with rules to search for concept candidates in Probase (Wu et al., 2012) and WordNet (Miller, 1995). The abstract knowledge triples connect conceptualized head events with their non-abstract counterparts from ATOMIC, forming commonsense knowledge at the concept level. Human annotations are used to verify the correctness of some conceptualizations and their resulting abstract commonsense triples. In total, 131K conceptualizations of 7K (45%) ATOMIC head events and 81K (1.3%) conceptualized triples are manually annotated, with a large number remaining unlabeled. The data is partitioned by following ATOMIC's original split of head events. Detailed statistics are shown in Table 9. In this paper, we evaluate all models using the test set from the annotated subset as the evaluation data. Meanwhile, we obtain CANDLE distilled models using the training set from the annotated subset to fine-tune discriminative models pre-trained on CANDLE conceptualizations. Supervised baselines are trained on the training set of AbstractATOMIC, while semi-supervised baselines also leverage the unlabeled data.

C.2 Generative Commonsense Inference

To evaluate COMET, we adopt the same evaluation setting employed by Hwang et al. (2021) for assessing commonsense generative models on the ATOMIC$^{20}_{20}$ dataset's test set. We use the entire test set, consisting of 34,689 triples across 23 different commonsense relations, to ensure the robustness of the evaluation. Additionally, we use the full training set to fine-tune models that were pre-trained on various CSKBs and CANDLE instantiations to fit them into the benchmark.

Recently, West et al. (2023) successfully trained a powerful commonsense inference generator using an open-format symbolic knowledge distillation framework. However, the dataset and models are not publicly released.

	aNLI	CSQA	PIQA	SIQA	WG
#QA Pairs	1,532	1,221	1,838	1,954	1,267
#Options	2	5	2	3	2

Table 10: Statistics on the number of QA pairs and the number of options for each question in benchmarks used in the zero-shot commonsense QA task.

C.3 Zero-shot Commonsense QA

We follow Ma et al. (2021); Wang et al. (2023a); Shi et al. (2023) and use the validation split of five commonsense QA benchmarks: Abductive NLI (aNLI; Bhagavatula et al., 2020), CommonsenseQA (CSQA; Talmor et al., 2019), PhysicalIQA (PIQA; Bisk et al., 2020), SocialIQA (SIQA; Sap et al., 2019b), and WinoGrande (WG; Sakaguchi et al., 2021). These benchmarks evaluate different aspects, including abductive reasoning, concept-level commonsense reasoning, physical commonsense understanding, emotional and social commonsense reasoning, and pronoun resolution. The validation splits are used as the official test sets may not be publicly available. Statistics on the number of QA pairs and the number of options per question are reported in Table 10.

D Implementation Details

This section provides additional implementation details in downstream task evaluations.

First, we use the Huggingface[3] Library (Wolf et al., 2020) to build all models. We reproduce all baselines according to implementation details described in their original papers. The reported results are consistent with their original papers if the same experiment is included. For CANDLE distilled models, please refer to the subsections (Appendix D.1, D.2, D.3) below.

For methods involving LLMs, we use their instruction fine-tuned versions as the backbone for the baselines. For LLAMA2, the accessed version is `meta-llama/Llama-2-7b/13b-chat-hf`. For Mistral, we use `mistralai/Mistral-7B-Instruct-v0.1`. This remains consistent whether we prompt them directly or fine-tune them for downstream tasks, as we have observed that the instruction-finetuned versions generally result in better performance. For ChatGPT,

[3] https://huggingface.co/

Task	Prompt
Event.	Given the event “PersonX enjoys drinking in the bar,” can “bar” be conceptualized as “entertainment venue”? Here, conceptualized means represented by a general concept. Answer ’Yes’ or ’No’ only without any other word.
Triple.	Given the assertion: PersonX enjoys drinking in entertainment venue, as a result, PersonX feels relaxed. entertainment venue is a general concept and represents many possible instances. Is this assertion plausible? Answer ’Yes’ or ’No’ only without any other word.
COMET	Please complete the given commonsense assertion with a few words. Don’t extend writing afterward. PersonX hears strange noises, as a result, PersonX will
aNLI	Premise: Jim decided to be a rockstar. Choice A: but didn’t know how to play an instrument. Jim signed up for guitar lessons. Choice B: Jim knew he would need to have a nickname. Jim signed up for guitar lessons. Which one is more likely to happen, given the premise? Only answer A or B without any other word.
CSQA	Question: He was at the gym trying to build muscle, what is it called that he is trying to build muscle on? Choice A: body of animal Choice B: arm Choice C: bodybuilder Choice D: body of dog Choice E: human body Which choice is correct? Only answer A or B or C or D or E without any other word.
PIQA	Goal: To remove an avocado from the shell Choice A: cut the avocado lengthwise, remove the pit, and scoop with a spoon Choice B: cut the avocado width wise, remove the pit, and scoop with a spoon Which choice can achieve the goal? Only answer A or B without any other word.
SIQA	Question: Robin went to the polls and posted her ballot for the candidate she wanted. As a result, Robin wanted to: Choice A: bomb the candidate Choice B: attend a rally Choice C: go home. Which choice is correct? Only answer A or B or C without any other word.
WG	Question: Jessica enjoyed a simple, basic life with Betty, but Choice A: Jessica was bored having a quiet existence. Choice B: Betty was bored having a quiet existence. Which choice is correct? Only answer A or B without any other word.

Table 11: Prompts used for evaluating LLM baselines across various tasks in a zero-shot scenario. Event. stands for event conceptualization discrimination and Triple. stands for triple conceptualization discrimination.

we access it through Microsoft Azure APIs[4]. The code of the accessed version for ChatGPT is `gpt-35-turbo-20230515`, and for GPT4 is `gpt-4-20230515`. The maximum generation length is set to 100 tokens for all tasks. For fine-tuning LLAMA2 and Mistral, we use the open code base of LLaMa-Factory[5]. Please refer to the subsections below for hyperparameter settings.

All experiments are conducted on sixteen NVIDIA-V100 (32G) GPUs.

For baselines involving prompting LLMs, we follow the approach done by Robinson and Wingate (2023) and Chan et al. (2024a), where each task is formulated in either a generative format or as multiple-choice QA. Table 11 shows the prompts used in zero-shot prompting scenarios. To incorporate five-shot exemplars, we include five randomly selected examples from the training set of each benchmark. These examples are merged into the prompt using the same format as the question, with the addition of including the answer at the end. For chain-of-thought reasoning, we prompt LLM in a two-step inference process. In the first step, we delve deeper into the question by requesting an intermediate-step rationale. Then, in the second step, we seek an answer based on the question and the previous step’s response by asking LLM to answer “Yes or No only” or select the correct option from a set of answers directly.

D.1 CSKB Conceptualization

For RoBERTa and DeBERTa-v3, we use a learning rate of 5e-6 and a batch size of 64. To optimize the models, we use an AdamW optimizer (Loshchilov and Hutter, 2019) and evaluate the model’s performance every 25 steps. The maximum sequence lengths for the tokenizers are set to 25 and 35

[4] https://azure.microsoft.com/en-us/products/ai-services/

[5] https://github.com/hiyouga/LLaMA-Factory

Training Data	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L	CIDEr	BERTScore
Backbone: GPT2-XL (Radford et al., 2019) *1.5B*								
Zero-shot	4.350	1.598	0.732	0.293	5.702	5.030	0.792	37.11
ATOMIC	32.23	19.06	13.27	10.28	17.63	25.50	20.15	58.39
+ Finetune	45.72	29.18	21.12	16.15	29.97	49.69	64.61	76.09
ATOMIC$^{20}_{20}$	42.15	25.77	17.82	13.14	29.82	47.61	63.70	70.39
ATOMIC-10X	33.69	18.82	11.71	7.910	18.78	25.69	19.29	61.47
+ Finetune	45.38	29.20	21.09	16.15	30.09	49.86	65.02	75.89
AbstractATOMIC	29.46	17.16	11.89	9.019	17.42	24.30	19.95	57.83
+ Finetune	45.30	29.08	21.00	16.06	29.98	48.61	63.98	75.56
CANDLE Distilled	26.91	16.44	12.31	10.28	17.66	23.66	21.36	57.15
+ Finetune	**50.71**	**33.85**	**25.55**	**20.43**	**32.45**	**51.91**	**69.68**	**76.86**
Backbone: ChatGPT (OpenAI, 2022) (`openai/gpt-3.5-turbo`)								
Zero-shot	11.82	4.258	1.891	0.926	13.87	13.73	4.350	49.28
Five-shot	26.32	12.50	7.160	4.415	18.60	24.65	8.313	58.69
Chain-of-thought	9.906	3.568	1.556	0.736	11.85	11.02	2.905	46.17
Backbone: LLAMA2 (Touvron et al., 2023) *7B*								
Zero-shot	18.26	7.453	3.594	1.945	15.90	20.28	8.872	48.23
Five-shot	31.22	16.87	9.767	5.989	19.74	27.67	17.83	58.41
ATOMIC	29.94	16.44	10.03	6.631	19.02	25.75	18.71	59.68
+ Finetune	42.04	23.01	14.10	9.125	27.80	42.90	53.17	71.52
ATOMIC$^{20}_{20}$	41.07	22.46	13.62	8.619	27.74	42.42	53.28	71.77
ATOMIC-10X	33.06	17.65	9.986	6.078	19.22	25.32	17.80	61.25
+ Finetune	42.07	23.08	14.14	9.198	28.14	42.75	53.69	71.93
AbstractATOMIC	26.08	13.27	7.799	5.018	15.08	21.20	14.78	56.83
+ Finetune	42.78	23.64	14.58	9.471	27.74	42.55	53.12	71.51
CANDLE Distilled	28.93	15.56	9.468	6.140	18.60	25.37	17.20	60.27
+ Finetune	43.86	24.40	15.12	10.00	28.36	43.86	54.25	72.94

Table 12: Full performances (%) of the commonsense inference modeling task (COMET) on the full test set of ATOMIC$^{20}_{20}$ (Hwang et al., 2021). The best performances using each backbone are underlined, and the best among all backbones are **bold-faced**. Finetune refers to fine-tuning back on the training set of ATOMIC$^{20}_{20}$.

for the two discriminative subtasks, respectively. Early stopping is used where the best checkpoint is selected when the largest validation accuracy is achieved. The models are trained on CANDLE distillation for one epoch and fine-tuned on the training set of AbstractATOMIC for one epoch.

For LLMs, such as LLAMA2 and Mistral, we use LoRA for fine-tuning, and the LoRA rank and α are set to 64 and 64. We use an Adam (Kingma and Ba, 2015) optimizer with a learning rate of 5e-6 and a batch size of 64. The models are fine-tuned for two epochs, and the checkpoint with the highest validation set accuracy is selected. All experiments are repeated three times using different random seeds, and the average performances are reported.

For VERA, we follow the exact same implementation[6] as released by Liu et al. (2023). To transform our binary classification subtasks into declarative formats, we begin by converting each piece of data into a declarative sentence using predefined natural language templates. Next, we create a corresponding negative statement by simply incorporating the word "not" into the correct sentence. For instance, a pair of statements is: "PersonX enjoys drinking at the bar. Bar is a social gathering place." and "PersonX enjoys drinking at the bar. Bar is not a social gathering place." The accessed backbone model is `liujch1998/vera`, and all other hyperparameter settings follow the default implementation. The model is trained on CANDLE distillation for one epoch and then fine-tuned on AbstractATOMIC for another.

D.2 Generative Commonsense Inference

For COMET, we implement the open-sourced code by Hwang et al. (2021) as our base to fine-tune the GPT2 model. The model is first pre-trained on CANDLE instantiations for one epoch, followed by fine-tuning on ATOMIC$^{20}_{20}$ for another epoch. An Adam (Kingma and Ba, 2015) optimizer is used with a learning rate of 1e-5 and a batch size of 32. A linear scheduler is used to decrease the learning rate gradually.

For LLAMA2-7B, we fine-tune it with the DeepSpeed framework (Aminabadi et al., 2022) by using FP16 as the precision. We optimize the model with an Adam (Kingma and Ba, 2015) optimizer with a learning rate of 1e-4 and a batch size of 64. The

[6] https://github.com/liujch1998/vera

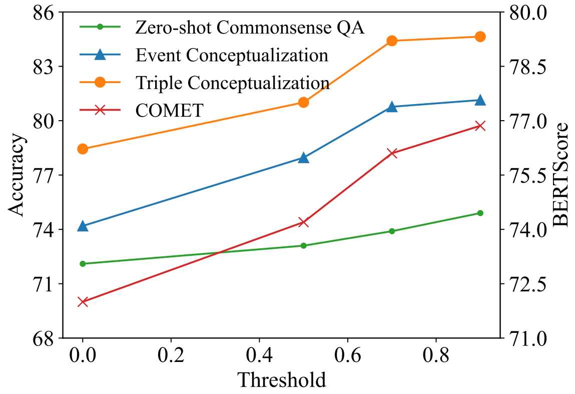

Figure 4: Ablation results examining the impact of different threshold values in CANDLE's critic filtering.

maximum length for the input and generated sentence concatenation is 500. We warm up the model with 3000 steps and evaluate the model every 1000 steps. A linear scheduler is also used. The LoRA rank is set to 8, and the α is set to 32.

In Table 12, we present supplementary automatic evaluation results, including models that have been pre-trained solely on CSKBs and CANDLE instantiations without subsequent fine-tuning.

D.3 Zero-shot Commonsense QA

For the task of zero-shot commonsense QA, we adopt the code base provided by Wang et al. (2023a)[7] and Liu et al. (2023)[8] to train two CANDLE distilled models. All hyperparameters and optimization strategies are kept unchanged from their original implementations as default settings. The models are trained for two epochs using QA pairs obtained from augmented-ATOMIC, including augmentations from ATOMIC-10X, AbstractATOMIC, and CANDLE instantiations.

Meanwhile, we present a comprehensive table presenting the results of all current methodologies for the task of zero-shot commonsense QA in Table 14. Notably, our CANDLE distilled models continue to exhibit strong performance compared to other models pre-trained on QA pairs sourced from multiple CSKBs. This serves as compelling evidence for the efficacy of CANDLE.

E Annotation Details

This paper utilizes expert annotations to assess the quality of distilled conceptualizations and instantiations, as well as evaluate the generations of different models for the COMET downstream task. We follow the annotation setting of Yu et al. (2023); Lu et al. (2024); Cheng et al. (2023), where four graduate students with ample experience in natural language processing research and expertise in commonsense reasoning are recruited as expert annotators to carry out the annotations. Their participation in the annotation process is voluntary and unpaid, in accordance with local laws, and is considered a contribution to this paper. Detailed instructions are provided to the annotators for each task, ensuring that they understand the requirements thoroughly. For CANDLE distillation evaluation, the annotators are asked to determine (1) the correctness of the conceptualizations and instantiations and (2) the plausibility of their formed triples. For COMET generation evaluation, they are asked to determine the plausibility of the generated triples. For each question, we also highlight the part to be considered by the annotators for their convenience. The annotation process is conducted independently, without any internal discussions among the annotators regarding the results. For each task, two annotators independently vote for each triple, and only when both annotators provide a positive vote will the triple be considered accepted or plausible. To prevent bias and ensure impartial results for CANDLE, the task input is randomly shuffled during the annotation process. As a result, the expert annotators achieve a pairwise agreement (IAA; Landis and Koch, 1977) of 0.80 and a Fleiss-kappa (Fleiss, 1971) of 0.61, indicating a remarkably high level of internal agreement.

F Ablation Study

In this section, we examine the impact of our critic filters on the ablation of CANDLE. Specifically, we investigate the effect of different levels of critic threshold or completely abandoning critic filtering on downstream tasks. We conduct four experiments with different settings, denoted as $t \in \{0, 0.5, 0.7, 0.9\}$, where $t = 0$ corresponds to abandoning critic filtering and using all distilled knowledge as complementary training data. For detailed statistics, please refer to Table 1. For each value of t, we select the distilled knowledge with a critic score higher than t and utilize it as complementary training data to train student models for the three downstream tasks. We employ the same training strategies described in the main body of the paper. In the case of CSKB conceptualization and zero-shot commonsense QA tasks, we utilize

[7] https://github.com/HKUST-KnowComp/CAR
[8] https://github.com/liujch1998/vera

Original	Concept./Instant.	Critic
PersonX swims in the lake, *as a result, PersonX feels*, tired.	PersonX swims in **freshwater**, *as a result, PersonX feels*, tired.	0.97
	PersonX swims in **the sea**, *as a result, PersonX feels*, tired.	0.87
	PersonX **swims**, *as a result, PersonX feels*, tired.	0.89
	PersonX **swims every week**, *as a result, PersonX feels*, tired.	0.81
PersonX is sitting in class, *as a result, PersonX will*, learns something.	PersonX is sitting in **instructional period**, *as a result, PersonX will*, learns something.	0.54
	PersonX is sitting in **a math class**, *as a result, PersonX will*, learns something.	0.75
	PersonX **study**, *as a result, PersonX will*, learns something.	0.78
	PersonX **learns how to do the exam**, *as a result, PersonX will*, learns something.	0.81
PersonX buys PersonY a gift, *as a result, PersonY feels*, joyful.	**remembrance**, *as a result, PersonY feels*, joyful.	0.19
	PersonX reminisce, *as a result, PersonY feels*, joyful.	0.27
	PersonX **shopping**, *as a result, PersonY feels*, joyful.	0.61
	PersonX **buys a new toy for PersonY**, *as a result, PersonY feels*, joyful.	0.90
PersonX always fought, *as a result, PersonY feels*, angry.	PersonX always **violent behavior**, *as a result, PersonY feels*, angry.	0.98
	PersonX always **punch others hardly**, *as a result, PersonY feels*, angry.	0.91
	combative personality, *as a result, PersonY feels*, angry.	0.98
	PersonX **PersonX likes to join a fight**, *as a result, PersonY feels*, angry.	0.85
PersonX gets a new bike, *as a result, PersonX wants*, to ride it.	PersonX gets a **transportation tool**, *as a result, PersonX wants*, to ride it.	0.92
	PersonX gets a **motor**, *as a result, PersonX wants*, to ride it.	0.98
	bike possession, *as a result, PersonX wants*, to ride it.	0.93
	PersonX has a nice bicycle, *as a result, PersonX wants*, to ride it.	0.89
PersonX spends time with PersonY, *PersonX is seen as*, social.	PersonX spends **love-building period** with PersonY, *PersonX is seen as*, social.	0.05
	PersonX spends **time in love** with PersonY, *PersonX is seen as*, social.	0.37
	social activity, *PersonX is seen as*, social.	0.64
	PersonX enjoys going to parties, *PersonX is seen as*, social.	0.73
PersonX hears sirens, *as a result, PersonX will*, make way to the siren.	**emergency response**, *as a result, PersonX will*, make way to the siren.	0.37
	PersonX sees an ambulance coming, *as a result, PersonX will*, make way to the siren.	0.74
	PersonX hears **loud noise**, *as a result, PersonX will*, make way to the siren.	0.67
	PersonX hears **a fire truck beeping**, *as a result, PersonX will*, make way to the siren.	0.77

Table 13: Case studies of **conceptualizations** and **instantiations** distilled from CANDLE in their original context. Original stands for the original triple sampled from ATOMIC. In the Concept./Instant. column, each box contains an abstract commonsense triple that includes **conceptualization**, followed by an instantiated commonsense triple with **instantiation**. We demonstrate two ways to conceptualize each original triple from ATOMIC.

DeBERTa-v3-large as the backbone model, with accuracy as the evaluation metric. For COMET, we use GPT2 and evaluate using the BERTScore as the evaluation metric. The results are visualized in Figure 4. Our analysis reveals a consistent trend where higher threshold values yield improved performance, indicating the reliability of our critic filter. However, it is worth noting that setting the threshold above 0.9 may potentially lead to even better performance. Nevertheless, such a trade-off comes with a downside: it reduces the amount of usable knowledge in each distillation round, which can impede the iterative process. The reason for this is that when the number of distilled conceptualizations and instantiations decreases significantly in each round, CANDLE is unable to incorporate new instantiated data for future distillation iterations. As a result, the "convergence" of those high-critic data occurs prematurely in CANDLE.

G Case Study

We present some examples in Table 13 to show conceptualizations and instantiations generated by CANDLE, along with their corresponding critic values assigned by our critic-filtering discriminators. It can be observed that both ChatGPT and LLAMA2 exhibit the ability to generate high-quality knowledge based on given instructions. Furthermore, they can introduce novel conceptualizations and events during the distillation chain, effectively meeting our expectations of CANDLE. Future works can investigate the feasibility of incorporating conceptualization and abstract knowledge into more downstream tasks, such as metaphysical reasoning (Wang and Song, 2024), complex reasoning (Fang et al., 2024; Bai et al., 2023; Ding et al., 2023), theory-of-mind reasoning (Chan et al., 2024b), text to table generation (Deng et al., 2024), and commonsense knowledge graph denoising (Deng et al., 2023).

Model/Method	CSKB	a-NLI	CSQA	PIQA	SIQA	WG	Avg.
Pre-trained Language Models							
Random Vote	-	50.0	20.0	50.0	33.3	50.0	40.7
Majority Vote	-	50.8	20.9	50.5	33.6	50.4	41.2
GPT2-L (Radford et al., 2019)	-	56.5	41.4	68.9	44.6	53.2	52.9
RoBERTa-L (Liu et al., 2019)	-	65.5	45.0	67.6	47.3	57.5	56.6
DeBERTa-v3-L (He et al., 2023)	-	59.9	25.4	44.8	47.8	50.3	45.6
Self-talk (Shwartz et al., 2020)	-	-	32.4	70.2	46.2	54.7	-
SMLM (Banerjee and Baral, 2020)	*	65.3	38.8	-	48.5	-	-
COMET-DynGen (Bosselut et al., 2021)	ATOMIC	-	-	-	50.1	-	-
MICO (Su et al., 2022)	ATOMIC	-	44.2	-	56.0	-	-
STL-PLM (Kim et al., 2022)	ATOMIC	71.6	64.0	72.2	63.2	60.5	66.3
MTL (Kim et al., 2022)	CWWV	69.6	67.3	72.5	52.0	57.2	63.7
MTL (Kim et al., 2022)	CSKG	69.8	67.1	72.0	61.9	59.3	66.0
STL-Adapter (Kim et al., 2022)	ATOMIC	71.3	66.5	71.1	64.4	60.3	66.7
STL-Adapter (Kim et al., 2022)	CSKG	71.5	66.7	72.1	64.7	59.0	66.8
RoBERTa-L (MR) (Ma et al., 2021)	ATM$_{10X}$	70.8	64.2	71.7	61.0	60.7	65.7
RoBERTa-L (MR) (Ma et al., 2021)	ATOMIC	70.8	64.2	72.1	63.1	59.2	65.9
RoBERTa-L (MR) (Ma et al., 2021)	CWWV	70.0	67.9	72.0	54.8	59.4	64.8
RoBERTa-L (MR) (Ma et al., 2021)	CSKG	70.5	67.4	72.4	63.2	60.9	66.8
DeBERTa-v3-L (MR) (Ma et al., 2021)	ATM10X	75.1	71.6	79.0	59.7	71.7	71.4
DeBERTa-v3-L (MR) (Ma et al., 2021)	ATOMIC	76.0	67.0	78.0	62.1	76.0	71.8
ZS-Fusion (Kim et al., 2022)	CWWV	69.6	67.6	73.1	53.7	59.5	64.7
ZS-Fusion (Kim et al., 2022)	CSKG	72.4	68.3	73.0	66.7	60.9	68.3
MKIF (Guan et al., 2023)	CSKG	72.5	71.0	73.1	-	61.0	-
CAR-RoBERTa-L (Wang et al., 2023a)	ATOMIC	72.3	64.8	73.2	64.8	61.3	67.3
CAR-RoBERTa-L (Wang et al., 2023a)	AbsATM	72.7	66.3	73.2	64.0	62.0	67.6
CAR-DeBERTa-v3-L (Wang et al., 2023a)	ATOMIC	78.9	67.2	78.6	63.8	78.1	73.3
CAR-DeBERTa-v3-L (Wang et al., 2023a)	AbsATM	79.6	69.3	78.6	64.0	78.2	73.9
DeBERTa-v3-L (CANDLE Distilled)	CANDLE	**81.2**$_{\uparrow 1.6}$	69.9$_{\uparrow 0.6}$	80.3$_{\uparrow 1.7}$	65.9$_{\uparrow 1.9}$	**78.3**$_{\uparrow 0.1}$	**74.9**$_{\uparrow 1.0}$
Large Language Models							
GPT-3.5 (`text-davinci-003`)	-	61.8	68.9	67.8	68.0	60.7	65.4
ChatGPT (`gpt-3.5-turbo`)	-	69.3	74.5	75.1	69.5	62.8	70.2
+ Chain-of-thought	-	70.5	75.5	79.2	**70.7**	63.6	71.9
+ Self-consistent chain-of-thought	-	73.2	**75.7**	81.7	69.7	64.1	72.9
GPT-4 (`gpt-4`)	-	75.0	43.0	73.0	57.0	77.0	65.0
LLAMA2 (7B; Touvron et al., 2023)	-	57.5	57.8	78.8	48.3	69.2	62.3
LLAMA2 (13B; Touvron et al., 2023)	-	55.9	67.3	80.2	50.3	72.8	65.3
Mistral-v0.1 (7B; Jiang et al., 2023)	-	51.0	59.6	**83.0**	42.9	75.3	62.4
VERA-T5-xxl (Liu et al., 2023)	ATOMIC	71.2	61.7	76.4	57.7	67.5	66.9
VERA-T5-xxl (Liu et al., 2023)	ATM10X	70.3	59.5	75.1	58.2	67.2	66.1
VERA-T5-xxl (Liu et al., 2023)	AbsATM	73.2	63.0	77.2	58.1	68.1	68.0
VERA-T5-xxl (CANDLE Distilled)	CANDLE	73.8$_{\uparrow 0.6}$	64.7$_{\uparrow 1.7}$	77.6$_{\uparrow 0.4}$	59.4$_{\uparrow 1.2}$	71.3$_{\uparrow 3.2}$	69.4$_{\uparrow 1.4}$
Supervised Learning & Human Performance							
RoBERTa-L (Supervised)	-	85.6	78.5	79.2	76.6	79.3	79.8
DeBERTa-v3-L (Supervised)	-	89.0	82.1	84.5	80.1	84.1	84.0
VERA-T5 (Multitask Supervised)	-	83.9	77.8	88.5	80.1	92.4	84.5
Human Performance	-	91.4	88.9	94.9	86.9	94.1	91.2

Table 14: Full zero-shot evaluation results (Accuracy%) on five commonsense question answering benchmarks. The best results are **bold-faced**, and the second-best ones are underlined. ↑ signifies the improvement CANDLE-distilled models achieve compared to the best baseline with the same backbone model. ATM10X stands for ATOMIC-10X (West et al., 2022) and AbsATM stands for AbstractATOMIC (He et al., 2024). All scores are retrieved from their original papers. For the GPT-X series, some results are retrieved from West et al. (2023).