CAT: A Contextualized Conceptualization and Instantiation Framework for Commonsense Reasoning

Weiqi Wang[1]*, Tianqing Fang[1]*, Baixuan Xu[1], Chun Yi Louis Bo[1], Yangqiu Song[1], Lei Chen[1,2]

[1]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China

[2]Information Hub, HKUST (GZ), Guangzhou, China

{wwangbw, tfangaa}@cse.ust.hk, bxuan@connect.ust.hk

{cybo, yqsong}@cse.ust.hk, leichen@ust.hk

Abstract

Commonsense reasoning, aiming at endowing machines with a human-like ability to make situational presumptions, is extremely challenging to generalize. For someone who barely knows about *meditation*, while is knowledgeable about *singing*, he can still infer that *meditation makes people relaxed* from the existing knowledge that *singing makes people relaxed* by first conceptualizing *singing* as a *relaxing event* and then instantiating that event to *meditation*. This process, known as conceptual induction and deduction, is fundamental to commonsense reasoning while lacking both labeled data and methodologies to enhance commonsense modeling. To fill such a research gap, we propose CAT (**C**ontextualized Conceptu**A**lization and Ins**T**antiation), a semi-supervised learning framework that integrates event conceptualization and instantiation to conceptualize commonsense knowledge bases at scale. Extensive experiments show that our framework achieves state-of-the-art performances on two conceptualization tasks, and the acquired abstract commonsense knowledge can significantly improve commonsense inference modeling. Our code, data, and fine-tuned models are publicly available at https://github.com/HKUST-KnowComp/CAT.

1 Introduction

"*Concepts are the glue that holds our mental world together.*"– Murphy (2004)

Commonsense reasoning is a crucial ability for machines to make situational presumptions and draw inferences from the knowledge that reflects our humans' understanding of situations and common facts (Davis, 1990; Davis and Marcus, 2015). It has gained increasing popularity in the Natural Language Processing (NLP) community with the emergence of CommonSense Knowledge Bases (CSKB) (Sap et al., 2019a; Speer et al., 2017; Hwang et al., 2021) and large language models (Bosselut et al., 2019; Rajani et al., 2019; Liu et al., 2022b; Su et al., 2022; Yu et al., 2022b). However, when encountering situations beyond the data given, more abstract background knowledge must be acquired and generalized to assist the reasoning (Tenenbaum et al., 2011), and language models trained with an autoregressive language modeling objective do not explicitly leverage such abstract knowledge during inference.

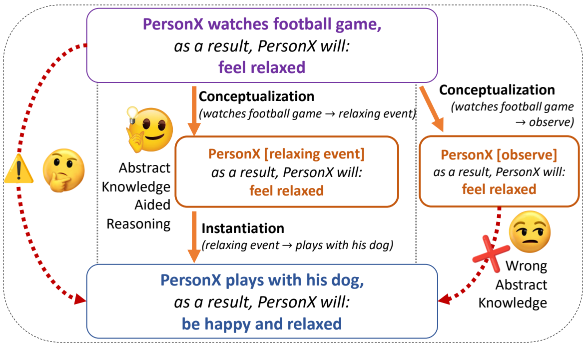

Figure 1: A demonstration of commonsense reasoning on an unknown situation, *PersonX plays with his dog*, with the aid of abstract commonsense knowledge. Decontextualized conceptualization, such as *observe*, may yield wrong abstract commonsense knowledge that cannot be instantiated within the corresponding context.

Instead, humans rely on conceptual induction and deduction (Murphy, 2004) to make inferences on novel situations without the need to memorize all special cases. As shown in Figure 1, humans can derive conceptualizations based on the assertion that "PersonX watches a football game, as a result, he feels relaxed" to infer that "relaxing events can make someone feel relaxed," where the acquired abstract commonsense knowledge can be further used as general knowledge to perform reasoning on similar or associated situations. A new commonsense knowledge "PersonX plays with his dog, as a result, he feels happy and relaxed" can be deduced by instantiating *relaxing events* to *playing with his dog*.

* Equal Contribution

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers, pages 13111–13140
July 9-14, 2023 ©2023 Association for Computational Linguistics

As the cornerstone of generalizable commonsense reasoning, such a process is extremely challenging for machines to replicate due to the absence of contextualized conceptualizations and abstract commonsense knowledge in CSKBs and a lack of relevant methodologies.

Yet, existing works address the process of induction and deduction separately via conceptualization and instantiation. Several methods performing conceptualization are proposed with a specific focus on entity-level (Durme et al., 2009; Song et al., 2011; Gong et al., 2016; He et al., 2020; Peng et al., 2022; Song et al., 2015) and event-level (Chen et al., 2020; He et al., 2022) semantics. Instantiation (Allaway et al., 2023), as the process that simulates conceptual deduction, is tackled separately and not leveraged by these methods. Though abstract commonsense knowledge can be derived by using existing conceptualization methods to abstract a certain instance from factual commonsense knowledge, several limitations still exist.

First, the plausibility of abstract commonsense knowledge banks on both the correctness of *conceptualization* and proper *contextualization* under specific assertions. The latter one, which is an essential step for the deduction of abstract knowledge, is missing from current methodologies. Take Figure 1 as an example, the concept *observe* will not necessarily lead to the result of "feeling relaxed", as *observe* omits the entertaining property of the original instance as a cost of abstraction. Second, instantiating abstract commonsense knowledge can yield much more and diverse concrete commonsense knowledge that can serve as an augmentation of the training dataset, while current methods undervalue such a process and only focus on conceptualization. Finally, the complex *contextualization* and *conceptualization* of commonsense knowledge can easily bring more than two orders of magnitude of data on top of the original dataset. This makes current labeled data scarce and infeasible for practitioners to annotate all of them, leaving a large amount of unlabeled data.

To fill in these research gaps, we propose CAT (**C**ontextualized Conceptu**A**lization and Ins**T**antiation), a semi-supervised learning framework that unites event conceptualization and instantiation in cascade to conceptualize CSKBs and acquire abstract commonsense knowledge to aid commonsense reasoning. Inspired by how humans learn with concepts (Carey, 2004), we design a novel bootstrapping[1] method to enhance conceptualizations and abstract commonsense knowledge verification with the help of similar conceptualizations and instantiations as a reference. We demonstrate the effectiveness of CAT by using the acquired abstract commonsense knowledge to train COMET (Bosselut et al., 2019), a commonsense inference language model that generates if-then commonsense knowledge, and showing that our derived abstract commonsense knowledge can significantly improve commonsense inference modeling.

Our contributions are three-fold: (1) We introduce a semi-supervised learning framework, CAT, to conceptualize CSKBs with the assistance of progressively bootstrapping similar abstract concepts or instantiations in the conceptualization process. (2) We use CAT to acquire abstract commonsense knowledge at scale with high quality, which can be used for commonsense inference modeling. (3) We demonstrate the effectiveness of our framework by achieving state-of-the-art performance on two CSKB conceptualization tasks and remarkably improving commonsense inference modeling with our derived abstract commonsense knowledge.

2 Related Works

Conceptualization and Instantiation. Many existing works have studied conceptualization and instantiation separately. Durme et al. (2009) first attempted to derive more general knowledge by abstracting over large sets of factoids obtained from WordNet (Miller, 1995) synsets. Song et al. (2011, 2015) and Gong et al. (2016) proposed to turn instances in a sentence into concepts via weight matching from Probase (Wu et al., 2012). Recently, Liu et al. (2022c) proposed a taxonomy-guided induction method to mine verb-oriented commonsense knowledge from verb phrases. Peng et al. (2022) constructed a conceptual knowledge benchmark to evaluate language models with three zero-shot probing tasks. While these works focus on the conceptualization of entities, He et al. (2022) constructed an event conceptualization benchmark based on ATOMIC (Sap et al., 2019a) by combining syntactic parsing, semantically heuristic matching, and human annotation. Besides, the line of works focusing on ultra-fine entity typing (Choi et al., 2018; Dai et al., 2021; Li et al., 2022) shared

[1]Bootstrapping refers to the linguistics term in language acquisition that humans learn new knowledge by recognizing its semantic elements and connecting them with known knowledge (Pinker and MacWhinney, 1987).

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

similar objectives of typing named entities, nominal nouns, and pronouns into a set of free-form phrases. Instantiation was attempted by Allaway et al. (2023), who proposed a controllable generative framework to probe valid instantiations for abstract knowledge automatically. Though Porada et al. (2021) and Peng et al. (2022) both proved that existing pretrained language models lack conceptual knowledge, none of existing works explicitly combine both techniques to derive abstract knowledge that is context-sensitive and generalizable.

Commonsense Reasoning. Endowing NLP systems with the ability to perform commonsense reasoning is an elusive goal of artificial intelligence (Sap et al., 2020). A diverse collection of commonsense reasoning tasks have been proposed as evaluation benchmarks (Talmor et al., 2019; Omura et al., 2020; Ponti et al., 2020; Fang et al., 2021a). Among them, Bosselut et al. (2019) proposed a generative model, COMET, to learn to produce *if-then* commonsense knowledge as an effective approach toward modeling commonsense inference that can be applied in various commonsense reasoning tasks (Talmor et al., 2019).

Semi-Supervised Learning. Semi-supervised learning (SSL) aims at taking advantage of unlabeled data to equip models with stronger generalization ability (van Engelen and Hoos, 2020). The most common approach is using pseudo labels (Iscen et al., 2019; Wang et al., 2022) to expose more unseen data to the student model. It has been applied in various machine learning tasks such as image classification (Liu et al., 2022a; Hu et al., 2021), text classification (Li et al., 2021; Meng et al., 2019; Xiao et al., 2019), commonsense knowledge base population (Fang et al., 2022), and named entity recognition (Liu et al., 2021; Chen et al., 2021).

3 Problem Definition

Definition. Conceptualizing an event-centric CSKB to derive abstract commonsense knowledge comprises two steps (He et al., 2022): event conceptualization and triple conceptualization.

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 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. As shown in Figure 1, conceptualizing *watch football game* to *observe* is wrong within the context of having *feel relaxed* as a result. Plausible (h_a, r, t) triples will be considered as valid abstract commonsense knowledge.

Specifically, in the first step, there are two ways of conceptualizing head events alone: a *retrieval-based discriminative* way and a *generative* way. The retrieval-based discriminative paradigm identifies and links a component i in h_o to a concept c in a concept taxonomy C to form a conceptualization h_a by replacing i with c. The model needs to verify whether h_a is a valid conceptualization of h_o. The generative paradigm aims to generate a h_a directly given h_o and the designated component i in h_o.

Formally, denote the annotated dataset in the first step, event conceptualization, as $D^l_h = \{(h_o, h_a, y)|h_o \in H_o, h_a \in H_a, y \in \{0, 1\}\}$, where h_o is an original head event without conceptualization, h_a is a corresponding conceptualization of h_o, and y is the human-annotated label indicating whether such a conceptualization is plausible or not. The labeled dataset in the second step, triple conceptualization, is denoted as $D^l_t = \{(h, r, t, y)|h \in H_a, r \in R, t \in T, y \in \{0, 1\}\}$, where h is a conceptualized head event from the first step, r and t are a relation and a tail from the original CSKB accompanied with the corresponding original head h_o, and y is the human-annotated label indicating whether such abstract commonsense knowledge, in the form of a conceptualized triple, is plausible or not. Besides labeled datasets, unlabeled datasets are defined similarly as D^u_h and D^u_t only with the difference that labels y are missing. Thus, the task objective for discriminative event conceptualization is to determine whether a h_o can be properly abstracted using h_a, where h_a is derived by replacing a component $i \subset h_o$ with

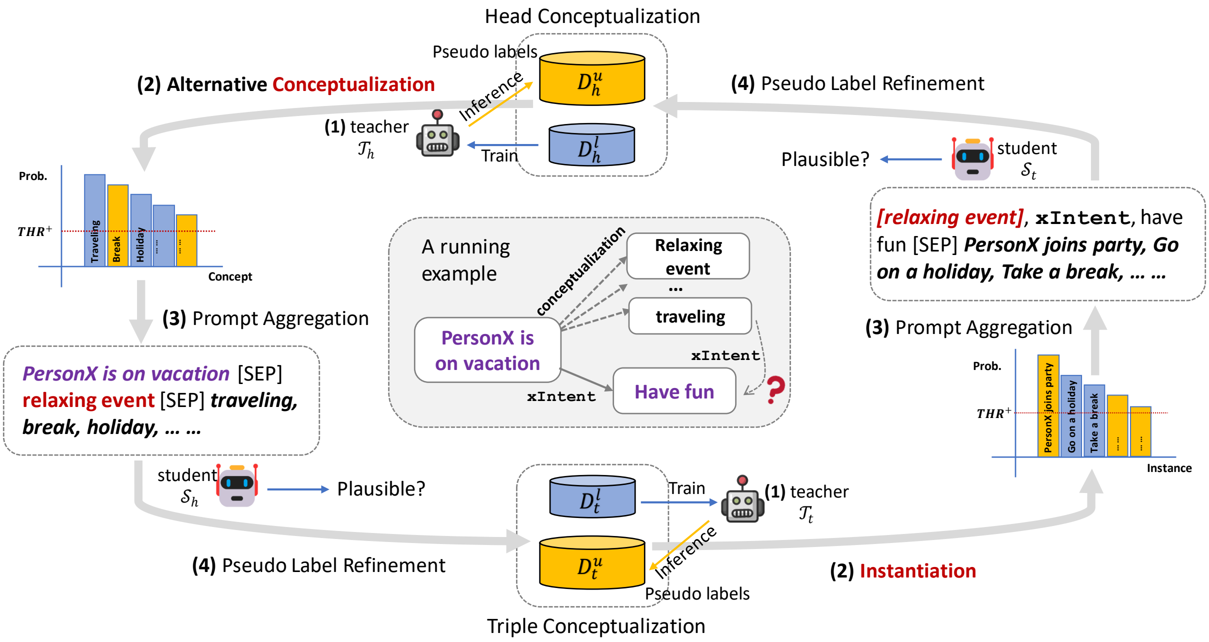

Figure 2: Overview of our CAT framework. A running example that conceptualizes the triple (PersonX is on vacation, `xIntent`, have fun) is presented in the figure, where the head is conceptualized first, and the model needs to determine whether the conceptualized triple still holds after the event conceptualization.

its linked concept c from a concept taxonomy C. The task objective for generative event conceptualization is to generate h_a directly from h_o with text generation models. For the triple conceptualization task, the objective is to distinguish whether a conceptualized triple (h_a, r, t), representing abstract commonsense knowledge, is plausible or not.

Dataset. To study conceptualization over CSKBs, we use the AbstractATOMIC dataset provided by He et al. (2022) as the benchmark. In AbstractATOMIC, ATOMIC is used as the original CSKB. And the event conceptualization adopts a *discriminative* way, where a syntactic parsing schema is defined to identify the components i in h_o to be heuristically linked to concept taxonomies Probase (Wu et al., 2012) and WordNet (Miller, 1995) to form conceptualized h_a. Such a heuristic can produce over 32 times more candidate conceptualized head events and over 10 times more conceptualized triples compared with the original ATOMIC, as the number of retrieved concepts from the concept taxonomy C can be manually controlled to acquire a large number of conceptualizations. Triple conceptualization is defined as predicting the plausibility of the triples whose head is conceptualized. Only 131K (26%) conceptualizations of 7K (45%) ATOMIC head events and 81K (1.3%) conceptualized triples are manually annotated as D_h^l and D_t^l, while others remain unlabeled D_h^u and D_t^u. The *trn/dev/tst* partition follows the same split as in the original ATOMIC. Statistics and more detailed explanations of AbstractATOMIC are shown in Table 1 and Appendix A.

4 CAT Framework

This section introduces our proposed Contextualized ConceptualizAtion and InsTantiation (CAT) framework for conceptualizing commonsense knowledge bases and acquiring abstract commonsense knowledge. An overview is presented in Figure 2. Our motivation is two-fold: first, adding instantiation after conceptualization to form a cycle can strongly benefit two conceptualization tasks simultaneously. On the one hand, instantiating conceptualized triple relies on the correctness of event conceptualization. On the other hand, properly conceptualized triples can benefit event conceptualization via instantiation by providing more context brought by (r, t). Second, to address the lack of annotations, we resort to pseudo labeling, a typical semi-supervised learning approach to automatically assign pseudo labels to the vast majority of unlabeled data using a teacher model.

Following He et al. (2022), we study the retrieval-based discriminative paradigm of event conceptualization and leave the generative paradigm as an intrinsic evaluation. In CAT,

we unify event conceptualization and triple conceptualization into one cycle and make them mutually benefit each other through instantiation and conceptualization. Our framework can be summarized into four steps:
(1) Train teacher models for both event conceptualization and triple conceptualization on the labeled dataset D^l_h and D^l_t, respectively. Use the two teachers to assign pseudo labels to unlabeled datasets.
(2) Conduct alternative conceptualization or instantiation on labeled and pseudo-labeled data.
(3) Bootstrap (aggregate) the alternative concepts and instances in the second step using natural language prompt templates and train student models on both labeled and pseudo-labeled data.
(4) Use the student models to refine the pseudo labels and then re-train the student models.

4.1 Teacher Model Training

Two teacher models on both event and triple conceptualization tasks are trained separately on the labeled dataset D^l_h and D^l_t. As both tasks are inherently text/triple classification, we adopt KG-BERT (Yao et al., 2019) as the skeleton of our models. The event conceptualization model determines whether h_a is a valid conceptualization of h_o, and the triple conceptualization model determines whether a conceptualized triple (h_a, r, t) is plausible or not. The two models θ are trained on annotated examples x_i with a cross-entropy loss (Eq. 1) and used to provide pseudo labels to instances from the unlabeled datasets D^u_h and D^u_t. Two thresholds, T^+ and T^-, are set to determine the pseudo labels of unlabeled examples with high confidence. Examples with a pseudo-labeled score higher than T^+ will be labeled $y_i = 1$, and those lower than T^- will be labeled $y_i = 0$. The rest will be discarded.

$$L(x_i, \theta) = -\sum_{i=1}^{|x|} y_i \log(\theta(x_i)) \quad (1)$$

4.2 Alternative Conceptualization and Instantiation

According to Murphy (2004), when humans learn a new concept, we pre-extract similar known concepts in our minds and infer possibly equivalent unknown concepts on the fly. Inspired by this theory, we retrieve additional abstract concepts or instantiated events to help discriminate conceptualizations and abstract commonsense knowledge. For event conceptualization, we retrieve some alternative possible conceptualizations of h_o to accompany the learning of h_a. Additional conceptualizations of h_o from both labeled and pseudo-labeled examples are predicted again by the teacher model and ranked according to their plausibility score prediction. And top m conceptualizations are retrieved with m being a hyperparameter to control the number of retrievals. For triple conceptualization, we perform instantiation in cascade to instantiate c to some concrete instances to assist the learning process. Possible instantiations of c are extracted from annotated and pseudo-labeled event conceptualizations by searching for conceptualized events $h'_a \in H_a$ other than h_a with c as the concept and extracting their corresponding instances $i \subset h'_a$. Similarly, the instances are then scored by the teacher model, and the top n of them are retrieved. Intuitively, alternative event conceptualizations can serve as hints for discriminating the correctness of the target conceptualization, and instantiations can carry additional contextualized information to help verify the plausibility of a conceptualized triple, which meets the objective of deriving abstract commonsense knowledge that is context-sensitive.

4.3 Prompt Aggregation

We then bootstrap the retrieved alternative conceptualizations/instantiations via natural language prompts. Here bootstrap (Carey, 2004) can be understood as binding the alternative retrievals and the target concept/triple together to strengthen the discrimination of the target concept/triple. As shown in Figure 2 step (3), the initially given input and retrieved concepts/instances are concatenated via human-defined prompts for both conceptualization tasks. Alternative concepts/instances are sorted in the order of their plausibility score ranking. Two student models $\mathcal{S}_h$ and $\mathcal{S}_t$ for both tasks are trained using the modified text with such prompts as inputs. They are expected to learn the bootstrapping connectionism between the target and the additional retrievals we provided. More detail about the prompt design is in Appendix B.

4.4 Pseudo-Label Refinement

All pseudo labels, initially derived by a teacher model trained on the original labeled dataset, are relabeled according to the plausibility score predicted by our newly enhanced student models $\mathcal{S}_h$ and $\mathcal{S}_t$. Similar to the teacher model, two thresholds, T^+ and T^-, are applied to distinguish positive and negative examples for both tasks. In addition, negative

Framework	Backbone PTLM / Method	Event Conceptualization		Triple Conceptualization	
		Validation	Testing	Validation	Testing
Supervised Learning	BERT-base *110M*	$82.4_{\pm0.05}$	$82.5_{\pm0.31}$	$71.2_{\pm0.58}$	$72.6_{\pm0.71}$
	BERT-large *340M*	$82.8_{\pm0.48}$	$83.1_{\pm0.80}$	$72.4_{\pm0.01}$	$73.7_{\pm0.00}$
	BART-base *139M*	$83.8_{\pm0.28}$	$84.4_{\pm0.32}$	$72.0_{\pm0.09}$	$72.6_{\pm0.15}$
	BART-large *406M*	$85.0_{\pm0.13}$	$85.2_{\pm0.22}$	$74.5_{\pm0.13}$	$76.2_{\pm0.19}$
	RoBERTa-base *110M*	$84.1_{\pm0.04}$	$84.5_{\pm0.19}$	$72.2_{\pm0.00}$	$74.1_{\pm0.00}$
	RoBERTa-large *340M*	$85.2_{\pm0.24}$	$85.5_{\pm0.02}$	$75.3_{\pm0.00}$	$76.9_{\pm0.01}$
	DeBERTa-v3-base *214M*	$85.1_{\pm0.08}$	$85.8_{\pm0.07}$	$73.9_{\pm0.10}$	$75.9_{\pm0.04}$
	DeBERTa-v3-large *435M*	$\underline{85.8}_{\pm0.05}$	$\underline{86.2}_{\pm0.15}$	$\underline{76.9}_{\pm0.03}$	$\underline{78.0}_{\pm0.02}$
	ELECTRA-base *110M*	$85.4_{\pm0.05}$	$85.8_{\pm0.02}$	$74.3_{\pm0.27}$	$76.2_{\pm0.12}$
	ELECTRA-large *340M*	$84.7_{\pm0.47}$	$85.3_{\pm0.38}$	$75.6_{\pm0.01}$	$77.9_{\pm0.06}$
	GPT2-base *117M*	$60.0_{\pm0.06}$	$59.1_{\pm0.14}$	$52.8_{\pm0.14}$	$55.9_{\pm0.11}$
	GPT2-medium *345M*	$61.2_{\pm0.11}$	$60.3_{\pm0.08}$	$54.6_{\pm0.17}$	$57.4_{\pm0.09}$
	GPT2-large *774M*	$64.1_{\pm0.05}$	$62.7_{\pm0.08}$	$60.5_{\pm0.11}$	$59.8_{\pm0.06}$
	GPT2-XL *1558M*	$64.2_{\pm0.19}$	$63.6_{\pm0.22}$	$62.2_{\pm0.08}$	$61.5_{\pm0.10}$
Semi-Supervised Learning	UDA (TF-IDF)	$83.6_{\pm0.29}$	$83.6_{\pm0.24}$	$75.8_{\pm1.26}$	$76.8_{\pm1.34}$
	UDA (back-trans.)	$83.4_{\pm0.27}$	$83.6_{\pm0.24}$	$75.8_{\pm1.25}$	$76.8_{\pm1.34}$
	Noisy-Student	$86.4_{\pm0.05}$	$86.5_{\pm0.09}$	$75.4_{\pm0.64}$	$76.7_{\pm0.59}$
	PseudoReasoner (BERT-base)	$83.3_{\pm0.11}$	$84.0_{\pm0.24}$	$73.0_{\pm0.14}$	$74.1_{\pm0.33}$
	PseudoReasoner (RoBERTa-large)	$\underline{86.6}_{\pm0.25}$	$\underline{86.7}_{\pm0.33}$	$\underline{76.3}_{\pm0.12}$	$\underline{77.2}_{\pm0.21}$
CAT ***(Semi-Supervised)***	BERT-base *110M*	$87.1_{\pm0.06}$	$87.4_{\pm0.11}$	$74.3_{\pm0.26}$	$76.3_{\pm0.38}$
	BERT-large *340M*	$87.7_{\pm0.16}$	$88.0_{\pm0.19}$	$75.8_{\pm0.23}$	$77.8_{\pm0.36}$
	BART-base *139M*	$88.2_{\pm0.09}$	$88.2_{\pm0.09}$	$75.7_{\pm0.09}$	$78.0_{\pm0.14}$
	BART-large *406M*	$88.6_{\pm0.07}$	$88.7_{\pm0.10}$	$77.2_{\pm0.12}$	$79.0_{\pm0.14}$
	RoBERTa-base *110M*	$88.4_{\pm0.12}$	$88.3_{\pm0.08}$	$76.9_{\pm0.16}$	$78.0_{\pm0.19}$
	RoBERTa-large *340M*	$89.0_{\pm0.15}$	$88.8_{\pm0.20}$	$78.2_{\pm0.08}$	$79.4_{\pm0.14}$
	DeBERTa-v3-base *214M*	$88.8_{\pm0.12}$	$88.9_{\pm0.08}$	$77.5_{\pm0.10}$	$79.9_{\pm0.07}$
	DeBERTa-v3-large *435M*	$\mathbf{89.1}_{\pm\mathbf{0.05}}$	$\mathbf{89.2}_{\pm\mathbf{0.14}}$	$\mathbf{78.7}_{\pm\mathbf{0.16}}$	$\mathbf{80.0}_{\pm\mathbf{0.33}}$
	ELECTRA-base *110M*	$88.7_{\pm0.10}$	$88.9_{\pm0.10}$	$74.9_{\pm0.15}$	$75.5_{\pm0.40}$
	ELECTRA-large *340M*	$88.6_{\pm0.77}$	$88.5_{\pm0.70}$	$74.9_{\pm0.15}$	$75.5_{\pm0.40}$

Table 2: Performance (%) by our CAT framework on the discriminative event conceptualization and triple conceptualization tasks. We report the average AUC score and standard deviation across experiments with three random seeds. The best performances within each framework are underlined, and the best among all models are bold-faced.

labels are assigned to triples whose conceptualized head events are predicted as wrong conceptualizations by $\mathcal{S}_h$, as wrong conceptualizations will not yield plausible abstract commonsense knowledge.

4.5 Application and Evaluation of CAT

The resulting models of CAT include an event conceptualization model and a triple conceptualization model, both fine-tuned on the refined pseudo labels and the labeled data. These two models can be used to conceptualize ATOMIC to a larger commonsense knowledge base on a more abstract level. We further conduct intrinsic evaluations on the acquired event conceptualization model under a generative event conceptualization paradigm and extrinsic evaluations on the resulting conceptualized CSKB with commonsense inference modeling task (COMET; Bosselut et al. (2019)) in Section 5. Here we select COMET as the representative because it is a general commonsense model that can be applied to various downstream commonsense reasoning tasks such as SocialIQA (Sap et al., 2019b), self-talk (Shwartz et al., 2020), and CSKB completion (Malaviya et al., 2020). Meanwhile, generative event conceptualization enables performing automatic conceptualization scalably. Both are important applications and evaluations of CAT.

5 Experiments

We conduct conceptualization experiments using CAT in Section 5.1 and generative experiments as evaluations in Section 5.2. These experiments demonstrate that CAT has a strong capability in conceptualizing CSKBs, and better conceptualization modeling can help populate more novel and diverse commonsense knowledge and thus help commonsense modeling (COMET).

5.1 CSKB Conceptualization

Baselines. We collectively introduce the baselines for both event and triple conceptualization tasks, as they are inherently classification tasks.

Training Data	BLEU-1		BLEU-2		METEOR		ROUGE-L		CIDEr		**Human**	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
$D^l_h + D^u_{0.95}$	**73.0**	<u>71.1</u>	**70.2**	63.0	**48.1**	**47.1**	**71.4**	<u>70.7</u>	**63.6**	<u>66.9</u>	**92.8**	**93.3**
$D^l_h + D^u_{0.9}$	<u>71.3</u>	**71.9**	65.2	<u>63.8</u>	<u>45.7</u>	<u>46.7</u>	<u>69.8</u>	**71.3**	<u>63.4</u>	**67.9**	<u>90.5</u>	<u>91.0</u>
$D^l_h + D^u_{0.8}$	68.2	68.4	<u>65.9</u>	**64.0**	44.8	44.0	66.6	66.7	60.0	62.0	86.0	85.7
$D^l_h + D^u_{0.7}$	66.5	67.2	57.2	62.6	43.0	43.4	65.9	65.8	60.4	61.2	79.0	80.3
$D^l_h + D^u_{0.5}$	64.9	62.4	58.3	51.1	41.2	40.9	63.8	63.0	58.2	59.4	74.5	79.0
D^l_h	67.6	65.3	56.8	53.1	43.5	43.1	65.7	66.6	60.2	60.9	70.0	81.5
Zero-Shot	20.2	17.0	6.80	4.11	5.80	4.70	3.80	3.00	1.90	1.60	15.0	11.5

Table 3: Performance (%) of GPT2 (XL) on the generative event conceptualization task. D^l_h stands for annotated labeled data, and D^u stands for the data acquired by CAT. The underfoot value indicates the threshold for selecting plausible pseudo labels. The best performances are bold-faced, and the second-best ones are underlined.

AUC is used as the evaluation metric. Under a supervised learning setting, we apply KG-BERT (Yao et al., 2019) model with BERT (Devlin et al., 2019), BART (Lewis et al., 2020), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021, 2023), and ELECTRA (Clark et al., 2020) as the backbone language models. We also attempt to leverage supervised generative language models as baselines. GPT2 (Radford et al., 2019) models are trained with a text generation objective only on positive examples, and we use perplexity as the prediction scores to calculate AUC. For the semi-supervised learning baselines, we leverage UDA (Xie et al., 2020a), NoisyStudent (Xie et al., 2020b), and PseudoReasoner (Fang et al., 2022) with RoBERTa-large being the backbone model. Additional explanations can be found in Appendix C.1.1.

Discriminative Results. The results for both tasks are presented in Table 2. Under a supervised learning setting, KG-BERT family mostly performs better on both tasks than GPT2 due to the fact that GPT2 is only fine-tuned on positive examples and thus cannot learn from negative examples that contain wrong conceptualizations and implausible abstract commonsense knowledge. As for the semi-supervised learning setting, previous SSL baselines are rather limited in improving the performance against supervised learning. The best PseudoReasoner only improves by 0.5% and 0.3% on the test set for both tasks compared with supervised RoBERTa-large models. Instead, models trained with CAT can outperform all other training methodologies. Comparing the test set performance with PseudoReasoner, small backbone models (BERT-base) can improve by 3.4% and 2.2%, and large models (RoBERTa-large) can be improved by 2.1% and 2.2%. This shows pipelining two-step conceptualizations as a loop and leveraging our proposed bootstrapping-based method can yield a larger performance gain compared with simply applying a semi-supervised learning strategy. Due to limited space, ablation studies on framework components and the semi-supervised learning paradigm of CAT are conducted in Appendix C.1.4. For example, the results indicate that bootstrapping alternative conceptualization and instantiation plays the most important role in assisting learning conceptualization among all components of CAT. Additional results and a computational cost study can be found in Appendix C.1.3 and Appendix D.

5.2 Application and Evaluation of CAT

As CAT is a framework for acquiring conceptualized commonsense knowledge, including both conceptualized head events (from h_o to h_a) and abstract commonsense triples (h_a, r, t), we assess these pseudo-labeled outcomes via two generative tasks with various threshold tuning as evaluations.

Generative Event Conceptualization. To intrinsically evaluate the effectiveness of CAT's event conceptualization, we use the acquired conceptualized head events as training data to learn a generative event conceptualizer. Specifically, the models are trained with instance-conceptualizations pairs in the format of "<*instance*> *is an instance of* <*concept*>". At the evaluation phase, the model is prompted with "<*instance*> *is an instance of* `[GEN]`" where <*instance*> is the instance to be conceptualized and `[GEN]` is the generation token. We then retrieve the top-1 generation and compare it against the target set from the evaluation dataset to compute four NLG metrics, as listed in Appendix C.2.1. These scores can be regarded as an approximation of the top-1 generations' recall.

Training Data	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Zero-Shot	5.42	4.89	1.84	1.51	0.65	0.52	0.26	0.21	6.50	5.70	6.40	5.90	1.60	1.20
ATOMIC (subset)	38.1	38.1	25.4	25.7	18.7	18.8	15.5	15.7	14.9	14.9	33.0	33.2	27.6	27.8
$+D^l_t$	38.1	38.5	24.8	25.5	17.8	18.4	14.7	15.2	15.3	15.6	33.1	33.7	26.8	27.3
+Finetune	38.6	39.0	25.8	26.6	18.9	19.7	15.7	16.4	15.1	15.4	33.6	34.4	28.8	30.0
$+D^u_{\text{Abs.ATM.}}$	40.0	40.3	27.1	27.8	20.0	20.8	16.5	17.5	16.1	16.3	35.3	35.7	31.6	31.7
+Finetune	40.1	40.5	27.1	27.8	20.1	20.8	16.7	17.4	16.2	16.4	35.4	35.9	31.8	31.7
$+D^l_t + D^u_{\text{Abs.ATM.}}$	40.2	40.6	26.2	27.4	19.0	20.4	15.1	16.8	16.3	16.5	35.0	35.4	31.0	31.3
+Finetune	40.0	40.4	26.0	26.9	18.7	19.7	15.0	16.1	16.3	16.4	35.0	35.4	30.3	30.7
$+D^u_{\text{CAT}}$	**41.2**	41.9	**28.1**	**29.0**	**20.7**	**21.5**	**16.5**	**17.8**	**16.6**	16.9	35.9	36.5	**33.4**	33.7
+Finetune	41.1	**42.0**	28.0	29.0	20.4	21.5	16.4	17.6	16.6	**17.0**	**36.0**	**36.8**	33.2	**33.8**
$+D^l_t + D^u_{\text{CAT}}$	39.9	40.5	26.2	27.4	19.3	20.6	16.0	17.4	16.0	16.2	35.0	35.4	30.8	31.3
+Finetune	40.4	41.0	26.6	27.6	19.5	20.7	16.1	17.1	16.2	16.5	35.4	35.8	31.3	31.5

Table 4: Performances (%) of GPT2 (XL) on commonsense inference modeling task (COMET). D^l_t stands for annotated abstract triples, and D^u_{CAT} stands for abstract triples acquired by CAT. $D^u_{\text{Abs.ATM.}}$ contains triples that are pseudo-labeled by a supervised RoBERTa discriminator, as done by He et al. (2022). The best performances are bold-faced. Finetune refers to fine-tuning back on the ATOMIC subset.

Additionally, we uniformly sample 500 generations from each evaluation split and conduct expert annotations on the plausibility of each conceptualization to ensure that out-of-domain concepts can be properly evaluated. The experts are asked to determine whether each top-1 generation is indeed a plausible conceptualization or not, such that the top-1 generations' precision is reflected. Thus, current evaluation measures jointly evaluate the top-1 generations' precision and recall, which makes it robust and non-easy to be impacted by repetition problems (Li et al., 2020). Zero-shot GPT2 and GPT2 fine-tuned on the originally labeled event conceptualizations in D^l_h are used as baselines. We also study the effect of the threshold T^+ that selects plausible conceptualized heads, where higher thresholds indicate higher plausibility regarded by CAT. The results are presented in Table 3. With a relatively high threshold, generators trained on a mixture of pseudo-labeled data by CAT and annotated concepts significantly outperform the baselines in every automated metric. A plausible rate of 93.3% is maximally achieved on the test set, which is 11.8% higher than the baseline. Gradually reducing the threshold also decreases the performance, indicating abstract heads with lower plausibility scores can be of poorer quality. Such results indicate that CAT can produce high-quality event conceptualizations for generative models to learn better conceptualizers without the need to annotate a large number of data.

Commonsense Inference Modeling (COMET). The second component of CAT produces triple-level abstract commonsense knowledge. We evaluate these abstract commonsense triples with a commonsense inference task that generates commonsense tails given heads and relations as inputs, as in COMET (Bosselut et al., 2019). Following He et al. (2022), we apply the same training and evaluation process to the models. The base training data we use are a subset of ATOMIC triples corresponding to those annotated abstract triples in D^l_t, which contains 17K (3.7%) among the original ATOMIC. We derive abstract commonsense knowledge using CAT from a subset of D^u_t where the heads correspond to those in the ATOMIC subset to ensure no data leakage, denoted as D^u_{CAT}. GPT2 is fine-tuned on the ATOMIC subset, the annotated abstract triples D^l_t, the abstract knowledge verified by CAT, or their combinations. The commonsense generation results are presented in Table 4. Similar to COMET (Bosselut et al., 2019), all models are evaluated on the original ATOMIC's full validation and testing sets. The best result is achieved using a mixture of the ATOMIC subset and abstract triples pseudo-labeled by our framework, with 0.95 as the threshold for selecting plausible triples. This indicates high-quality abstract commonsense triples can indeed provide a more general view of the original commonsense knowledge, thus helping commonsense inference. Additionally, training with our pseudo-labeled examples outperforms training with those annotated triples in AbstractATOMIC, which also validates the effectiveness of our model that leverages a large amount of unlabeled data. To further investigate how conceptual knowledge

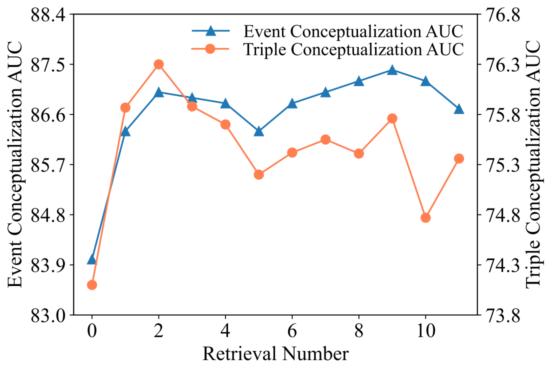

Figure 3: Ablation study on the number of retrieved conceptualizations/instantiations for CAT framework.

improves commonsense inference modeling, we conduct more empirical analysis in Section 5.4. Additional experiment results with other thresholds and case studies can be found in Appendix C.2.3 and Appendix E, respectively.

5.3 Number of Retrieved Alternative Conceptualizations and Instantiations.

We then study the ablation of bootstrapping different numbers of alternative conceptualizations/instantiations (denoted as #retrieval) in our CAT framework. For simplicity, when tuning the #retrieval for one task, the #retrieval of the other task is fixed at the best value we acquired. We plot the test AUC score with #retrieval from 0 to 11 using BERT-base as the backbone model in Figure 3. #retrieval=0 refers to training with a simple student-teacher framework without bootstrapping alternative conceptualizations and instantiations. For event conceptualization, the performance generally positively correlates with the number of retrievals, while it starts dropping after 9. A reversed trend is observed for triple conceptualization, where using only two instances achieves the best performance. One possible reason is that in triple conceptualization, the retrieved instances are events and much longer than the retrieved concepts in event conceptualization, and aggregating various alternative events for a triple will cause language models to be less sensitive to the semantics of the original triple (Holtzman et al., 2020).

5.4 The Effect of Abstract Knowledge

We finally study the effect of abstract commonsense knowledge acquired by CAT by studying the semantic overlaps between training and testing data. We sort the test set by the BERTScore (Zhang et al., 2020b) between each individual testing entry against the whole training set in the original ATOMIC and split them in half to acquire two test groups. The testing entries with lower BERTScore on the training set indicate a larger semantic shift from the training set (Deutsch and Roth, 2021), which is also harder for models to discriminate (Hsu et al., 2020). We denote the testing group with a lower BERTScore as "Difficult" and the other half as "Easy". The performance gain on the two test set splits between the best conceptualization-aided COMET and the COMET trained on the ATOMIC subset only is reported in Figure 4. We can observe that training COMET with abstract commonsense knowledge leads to a larger improvement for harder test examples dissimilar from the original training set, indicating that introducing extra abstract commonsense knowledge can help COMET become more generalizable to harder test sets.

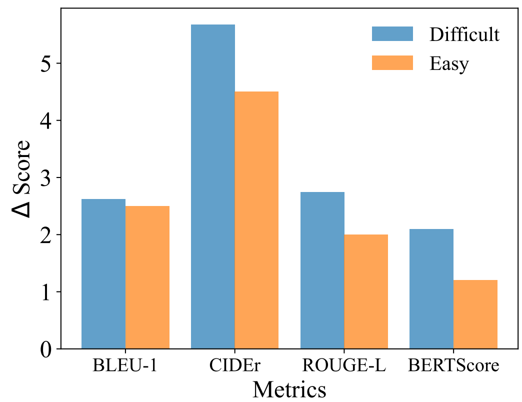

Figure 4: Comparison of performance improvement by GPT2 generator trained on the conceptualization-aided ATOMIC subset for two groups of testing head events.

6 Conclusion

In conclusion, this paper proposes CAT, a semi-supervised learning framework for commonsense reasoning, by leveraging the power of abstract commonsense knowledge. By achieving state-of-the-art performances in CSKB conceptualization tasks, we remarkably improve modeling commonsense inference, as an important cornerstone of many commonsense reasoning tasks. Our analysis also demonstrates that high-quality abstract commonsense knowledge can benefit commonsense inference modeling by providing more generalizability on hard commonsense knowledge. We hope this work can draw insights toward commonsense reasoning from a conceptualization perspective.

Limitations

Our framework manually sets thresholds T^+ and T^- in pseudo labeling by observations of data quality and hyperparameter searching. Dynamic threshold tuning (Xu et al., 2021) or meta pseudo labels (Pham et al., 2021; Li et al., 2021) can be implemented to better filter pseudo-labeled examples. And the thresholds for different tasks can be tuned separately to improve the models' generalizability.

Recently, large generative language models such as GPT3.5 (Brown et al., 2020) and ChatGPT[2] (Ouyang et al., 2022; Gao et al., 2022) have demonstrated their strong potential on various NLP tasks including probing abstract commonsense knowledge with in-context learning (Brown et al., 2020; Xie et al., 2022). Due to our limited access, we did not conduct fully-scaled experiments in our paper. A short discussion with case studies is provided in Appendix E.3.

While our framework only operates on AbstractATOMIC as the conceptualization of ATOMIC, it's also worthy of verifying our framework on other CSKBs such as ATOMIC2020 (Hwang et al., 2021), GLUCOSE (Mostafazadeh et al., 2020), ATOMIC10X (West et al., 2022), FolkScope (Yu et al., 2022a) and eventuality CSKB such as ASER (Zhang et al., 2020a, 2022) and constructing large conceptualized CSKB benchmarks. In addition, we only evaluated the power of the acquired abstract commonsense knowledge on the commonsense knowledge generation task (COMET), while other commonsense reasoning tasks remain future works (Wang et al., 2023a), such as COLA (Wang et al., 2023b), CommonsenseQA (Talmor et al., 2019, 2021), SocialIQA (Sap et al., 2019b), Winograd Schema Challenge (Levesque et al., 2012), PIQA (Bisk et al., 2020), Abductive Commonsense Reasoning (Bhagavatula et al., 2020), and Winogrande (Sakaguchi et al., 2020).

Ethics Statement

This paper introduces CAT, a framework for commonsense reasoning via conceptualizing CSKB to acquire abstract commonsense knowledge. The experiments are conducted on publicly available and well-established datasets that are shared via open-access licenses. The usage of these datasets in our paper is only for research purposes and is consistent with the datasets' intended usage. The primary dataset, AbstractATOMIC, largely shares the content with another CSKB, ATOMIC, which is anonymized and desensitized (Sap et al., 2019a). Thus, no data privacy issue is involved.

The potential risks of CAT are relatively low. Since CAT is trained on AbstractATOMIC, a conceptualization benchmark based on a popular CSKB, ATOMIC, and two concept taxonomies, Proabse and WordNet, it is expected that CAT does not contain any private, offensive, biased, and sensitive information or social, political issues. The studied tasks all focus on conceptualization or CSKB, which is not likely to generate harmful content, as shown in the case studies in Appendix E. Thus, we believe that CAT does not yield additional risks.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive comments. The authors of this paper are 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, the MHKJFS (MHP/001/19) from ITC of Hong Kong and the National Key R&D Program of China (2019YFE0198200) with special thanks to HKMAAC and CUSBLT. We also thank the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21, RMGS23CR05, RMGS23EG08).

References

Emily Allaway, Jena D. Hwang, Chandra Bhagavatula, Kathleen 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*, pages 2610–2627, Dubrovnik, Croatia. Association for Computational Linguistics.

Mostafa M. Amin, Erik Cambria, and Björn W. Schuller. 2023. Will affective computing emerge from foundation models and general ai? A first evaluation on chatgpt. *CoRR*, abs/2303.03186.

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.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable

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

but inexperienced solver: An investigation of commonsense problem in large language models. *CoRR*, abs/2303.16421.

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, 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.

Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.*, 30(7):1145–1159.

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*.

Susan Carey. 2004. Bootstrapping & the origin of concepts. *Daedalus*, 133(1):59–68.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

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.

Weile Chen, Huiqiang Jiang, Qianhui Wu, Börje Karlsson, and Yi Guan. 2021. Advpicker: Effectively leveraging unlabeled data via adversarial discriminator for cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 743–753. Association for Computational Linguistics.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 87–96. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hongliang Dai, Yangqiu Song, and Haixun Wang. 2021. Ultra-fine entity typing with weak supervision from a masked language model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1790–1799. Association for Computational Linguistics.

Ernest Davis. 1990. *Representations of commonsense knowledge*. notThenot Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning, CoNLL 2021, Online, November 10-11, 2021*, pages 300–309. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Li Du, Xiao Ding, Ting Liu, and Zhongyang Li. 2019. Modeling event background for if-then commonsense reasoning using context-aware variational autoencoder. 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 2682–2691. Association for Computational Linguistics.

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, Quyet V. Do, Sehyun Choi, Weiqi Wang, and Yangqiu Song. 2023. Ckbp v2: An expert-annotated evaluation set for commonsense knowledge base population. *CoRR*, abs/2304.10392.

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.

Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization. *CoRR*, abs/2210.10760.

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.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction, AKBC@CIKM 13, San Francisco, California, USA, October 27-28, 2013*, pages 25–30. ACM.

Masato Hagiwara, Yasuhiro Ogawa, and Katsuhiko Toyama. 2006. Selection of effective contextual information for automatic synonym acquisition. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*. The Association for Computer Linguistics.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2022. Acquiring and modelling abstract commonsense knowledge via conceptualization. *CoRR*, abs/2206.01532.

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, ICLR 2023*. OpenReview.net.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. 2020. Generalized ODIN: detecting out-of-distribution image without learning from out-of-distribution data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10948–10957. Computer Vision Foundation / IEEE.

Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. 2021. Simple: Similar pseudo label exploitation for semi-supervised classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 15099–15108. Computer Vision Foundation / IEEE.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: BERT for word sense disambiguation with gloss knowledge. 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 3507–3512. Association for Computational Linguistics.

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.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. 2019. Label propagation for deep semi-supervised learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5070–5079. Computer Vision Foundation / IEEE.

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*.

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.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10-14, 2012*. AAAI Press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Bangzheng Li, Wenpeng Yin, and Muhao Chen. 2022. Ultra-fine entity typing with indirect supervision from natural language inference. *Trans. Assoc. Comput. Linguistics*, 10:607–622.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4715–4728. Association for Computational Linguistics.

Zheng Li, Danqing Zhang, Tianyu Cao, Ying Wei, Yiwei Song, and Bing Yin. 2021. Metats: Meta teacher-student network for multilingual sequence labeling with minimal supervision. 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 3183–3196. 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.

Fengbei Liu, Yu Tian, Yuanhong Chen, Yuyuan Liu, Vasileios Belagiannis, and Gustavo Carneiro. 2022a. ACPL: anti-curriculum pseudo-labelling for semi-supervised medical image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20665–20674. IEEE.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022b. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3154–3169. Association for Computational Linguistics.

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

Kun Liu, Yao Fu, Chuanqi Tan, Mosha Chen, Ningyu Zhang, Songfang Huang, and Sheng Gao. 2021. Noisy-labeled NER with confidence estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3437–3445. Association for Computational Linguistics.

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.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. 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 2925–2933. AAAI Press.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-supervised hierarchical text classification. 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 6826–6833. AAAI Press.

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

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David W. Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: generalized and contextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4569–4586. Association for Computational Linguistics.

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

Kazumasa Omura, Daisuke Kawahara, and Sadao Kurohashi. 2020. A method for building a commonsense inference dataset based on basic events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2450–2460. Association for Computational Linguistics.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

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. 2022. 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.

Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V. Le. 2021. Meta pseudo labels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11557–11568. Computer Vision Foundation / IEEE.

Steven Pinker and B MacWhinney. 1987. The bootstrapping problem in language acquisition. *Mechanisms of language acquisition*, pages 399–441.

Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2362–2376. Association for Computational Linguistics.

Ian Porada, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2021. Modeling event plausibility with consistent conceptual abstraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1732–1743. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *CoRR*, abs/2302.06476.

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.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. 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 4932–4942. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. 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 8732–8740. AAAI Press.

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.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, ACL 2020, Online, July 5, 2020*, pages 27–33. 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*, pages 1339–1351, Abu Dhabi, United Arab Emirates. 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.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. Commonsenseqa 2.0: Exposing the limits of AI through gamification. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

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.

Jesper E. van Engelen and Holger H. Hoos. 2020. A survey on semi-supervised learning. *Mach. Learn.*, 109(2):373–440.

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.

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: generative pseudo labeling for unsupervised domain adaptation of dense retrieval. 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 2345–2360. Association for Computational Linguistics.

Peifeng Wang, Filip Ilievski, Muhao Chen, and Xiang Ren. 2021. Do language models perform generalizable commonsense inference? In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3681–3688. Association for Computational Linguistics.

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. *CoRR*, abs/2305.14869.

Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023b. COLA: contextualized commonsense causal reasoning from the causal inference perspective. *CoRR*, abs/2305.05191.

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.

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.

Huiru Xiao, Xin Liu, and Yangqiu Song. 2019. Efficient path prediction for semi-supervised and weakly supervised hierarchical text classification. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3370–3376. ACM.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020a. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Qizhe Xie, Minh-Thang Luong, Eduard H. Hovy, and Quoc V. Le. 2020b. Self-training with noisy student improves imagenet classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10684–10695. Computer Vision Foundation / IEEE.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. 2021. Dash: Semi-supervised learning with dynamic thresholding. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11525–11536. PMLR.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. KG-BERT: BERT for knowledge graph completion. *CoRR*, abs/1909.03193.

Changlong Yu, Jialong Han, Peifeng Wang, Yangqiu Song, Hongming Zhang, Wilfred Ng, and Shuming Shi. 2020. When hearst is not enough: Improving hypernymy detection from corpus with distributional models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6208–6217. Association for Computational Linguistics.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2022a. Folkscope: Intention knowledge graph construction for discovering e-commerce commonsense. *CoRR*, abs/2211.08316.

Changlong Yu, Hongming Zhang, Yangqiu Song, and Wilfred Ng. 2022b. Cocolm: Complex commonsense enhanced language model with discourse relations. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1175–1187. Association for Computational Linguistics.

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.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020a. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.

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

Appendices

A Dataset Description

In this section, we introduce more about AbstractATOMIC (He et al., 2022), as the primary dataset we experimented with. AbstractATOMIC is a conceptualized commonsense knowledge benchmark that is built upon ATOMIC (Sap et al., 2019a), a popular CSKB in the format of (h, r, t) triples. The dataset is entirely in English. It contains two parts of data: (1) event conceptualization data and (2) abstract knowledge triples conceptualization data.

The event conceptualization data contain conceptualizations for head event instances, where the events are filtered from the original ATOMIC head events. Unlike the traditional entity concept taxonomies, where instances are nouns or verb phrases, AbstractATOMIC includes instance candidates that can be either the entire head event or a certain component of an event. Detailed examples can be found in Appendix E.

The instances within each head event are identified through syntactic parsing by using a parser from the spaCy [3] library and matching with five human-defined rules. After identification, the candidate instances will be heuristically matched against Probase (Wu et al., 2012) and WordNet (Miller, 1995) via GlossBERT (Huang et al., 2019) to acquire their candidate concepts. A neural generator based on GPT2, similar to the baseline in this paper, is also trained to generate concepts. A supervised conceptualization verifier, based on RoBERTa (Liu et al., 2019), is trained as the final gatekeeper to verify the acquired concepts roughly.

	D^l_h	D^u_h	Total
#Unq. event	7,196	15,165	15,388
#Unq. instance	7,935	20,843	21,493
#Unq. concept	20,036	20,367	31,227
Avg. #concept/event	18.21	24.57	32.73
Avg. #concept/instance	16.51	17.88	23.43

Table 5: Additional statistics of the event conceptualization data in AbstractATOMIC (AbsATM). D^l_h stands for annotated event conceptualizations and D^u_h are unverified conceptualizations. # denotes "number of", Unq stands for unique, and Avg is average.

Human annotations on the Amazon Mechanical Turk platform are further conducted to acquire annotations on the correctness of 131K conceptualizations of 7K ATOMIC events. All conceptualizations that are not annotated are regarded as unlabeled data in this paper. More detailed statistics for the head event conceptualization data can be found in Table 5.

After acquiring the event conceptualizations by only focusing on head events, abstract commonsense knowledge, in the form of (h, r, t) triple, is collected by connecting conceptualized head event with its non-abstract counterparts (commonsense relations and inference tails) from ATOMIC. Only the head events contain abstract concepts. Thus, these abstract triples are more generalized if-then commonsense knowledge that is potentially useful for commonsense reasoning through instantiation.

Human annotations on Amazon Mechanical Turk further verify 81K uniformly sampled abstract triples. These triples only correspond to 689 unique ATOMIC head events, which makes annotations relatively scarce compared with the scale of unlabeled data. A supervised RoBERTa-large verifier is trained on the annotated triples to roughly verify abstract triples that are not annotated. Triples with scores higher than 0.9 are pseudo-labeled as positive ones (He et al., 2022). However, this paper only leverages these pseudo-labeled examples in the commonsense inference generation task (COMET) as baselines. Only annotated triples are considered hard-labeled for all other tasks concerned. And triples that are not annotated are treated as unlabeled by default. The detailed relational distribution of abstract triples is presented in Table 6. Examples can be found in Appendix E.

Relation	ATOMIC	D^l_t	D^u_t	$D^u_{\text{Abs.ATM.}}$
xEffect	78,832	12,168	938,330	451,564
oEffect	28,351	3,526	333,845	160,207
xWant	101,249	15,312	1,170,835	543,964
oWant	43,079	5,408	484,570	227,493
xReact	62,969	8,923	510,476	288,019
oReact	26,570	3,030	224,706	126,386
xNeed	74,272	11,733	900,429	425,060
xAttr	110,791	14,249	838,191	465,511
xIntent	45,490	6,848	519,813	259,694
Total	572,053	81,197	5,921,195	2,947,898

Table 6: Abstract commonsense triple distribution by relations. D^l_t stands for annotated triples and D^u_t are unverified triples. $D^u_{\text{Abs.ATM.}}$ stands for abstract triples verified by a supervised RoBERTa-large discriminator, as done by He et al. (2022).

[3] https://spacy.io/

B Prompt Design

In this section, we introduce the textual prompts used for training various models.

For event conceptualization, denotes the original event as h_o, instance as i, target concept to be verified as c, and retrieved alternative conceptualizations as $c_{r,1}, c_{r,2}, c_{r,3}, ..., c_{r,m}$. The prompt for training the teacher model is "`[CLS]` h_o `[SEP]` c", while the one for training the student model is "`[CLS]` h_o `[SEP]` c `[SEP]` $c_{r,1}$, $c_{r,2}$, $c_{r,3}$, ..., $c_{r,m}$". For the example in Figure 2, the filled prompt is "PersonX is on vacation `[SEP]` relaxing event `[SEP]` traveling, break, holiday." Specifically, special tokens <c> and </c> are used to enclose $i \subset h_o$ within the original event to highlight the instance to be conceptualized. GPT2 generators use similar prompts, with the difference that `[SOS]` and `[EOS]` special tokens are inserted to denote the start and end of the sentence, respectively.

For triple conceptualization, denotes the head, relation, and tail of an abstract commonsense triple as (h, r, t), the abstract concept in the conceptualized head as $c \subset h$, and retrieved instantiations as $e_{r,1}, e_{r,2}, e_{r,3}, ..., e_{r,n}$. The prompt for training generally follows the one used by He et al. (2022). For the teacher model, "`[CLS]`, h_1, ..., $h_{|h|}$, `[SEP]`, $[r]$, `[SEP]`, t_1, ..., $t_{|t|}$" is used as the prompt. Similarly, student models are trained with a prompt "`[CLS]`, h_1, ..., $h_{|h|}$ `[SEP]` $[r]$ `[SEP]` t_1, ..., $t_{|t|}$ `[SEP]` $e_{r,1}$, $e_{r,2}$, $e_{r,3}$, ..., $e_{r,n}$". A filled example by using the case in Figure 2 is "relaxing event `[SEP]` because PersonX wanted `[SEP]` have fun `[SEP]` PersonX joins party, go on a holiday, Take a break." The commonsense relation within each triple is translated into human-readable text, as shown in Table 7.

Relation	Human Readable Text
xEffect	as a result, PersonX will
oEffect	as a result, PersonY or others will
xWant	as a result, PersonX want
oWant	as a result, PersonY or others want
xReact	as a result, PersonX feel
oReact	as a result, PersonY or others feel
xIntent	because PersonX wanted
xNeed	before that, PersonX needed
xAttr	PersonX is described as

Table 7: Textual prompt for commonsense relations (Fang et al., 2021b). Commonsense triple (h, r, t) is translated to human language *"if h, [prompt] t"*.

The generative event conceptualization by GPT2 generators uses "`[SOS]` h_o `[SEP]` i `[GEN]`" as the input template, where `[GEN]` indicates the special token for generation. Commonsense inference modeling uses the same prompt as done by Hwang et al. (2021); Fang et al. (2021b).

In addition, we observe that adding special tokens such as <c> and </c> can effectively boost performance. But adding textual guidelines such as "is an instance of" or "is a concept of" does not have any positive effect. The same trend is observed for the bootstrapping prompt, where adding external texts such as "is also instances of" or "can be instantiated to" will harm the model significantly.

C Additional Experiments

In this section, we present additional details and experiment results for CSKB conceptualization tasks (Appendix C.1) and applications, as well as evaluations, of CAT (Appendix C.2) that are not covered in the paper due to limited space.

C.1 CSKB Conceptualization

C.1.1 Baselines

For supervised learning baselines of both discriminative conceptualization tasks, KG-BERT (Yao et al., 2019) is adapted as the skeleton of our baseline models. For BART, we use the embedding of the end-of-sentence token in the decoder as the representation of the input sequence. For other models, the embedding of the [CLS] token is used as the representation vector. Linear layers are appropriately appended after the encoder model to perform text classification.

For the semi-supervised baselines, we provide additional explanations for different methods:

UDA. In the original paper of UDA (Xie et al., 2020a), two data augmentation methods, back-translation and TF-IDF replacement, are implemented for unsupervised data augmentation. We leverage both methods in our conceptualization tasks as two different baselines. For the triple conceptualization task, we follow the same setting as proposed in PseudoReasoner (Fang et al., 2022). The back-translation method translates the original corpus from English to French and then translates it back. Special replacements are taken to avoid the influence of special tokens. Meanwhile, the TF-IDF method uses a probability of 0.1 to replace the original corpus according to its TF-IDF score. For the event conceptualization task, we concatenate the head event and its annotated concept into one new sentence and then feed it into the model. For the unlabeled conceptualizations, we enclose the instance and concept with special tokens <c> and </c>, which is the same as our framework, and then use back translation or TF-IDF to generate the augmented data. The input for triple conceptualization follows a similar way as supervised baselines. It is observed that these special tokens will not affect the translation significantly as they will be preserved in the translation output. Last but not least, the model θ is trained on a mixture of annotated data x_1 and augmented data x_2 by using the consistency training loss, as shown in Equation 2.

$$J(\theta) = \mathbb{E}_{x1 \sim P_L(x)}[-\log p_\theta(y_1|x_1)] + \lambda \mathbb{E}_{x2 \sim P_U(x)} \mathbb{E}_{\hat{x} \sim q(\hat{x}|x_2)})[CE(p_{\tilde{\theta}}(y|x_2)||p_\theta(y|\hat{x})] \tag{2}$$

NoisyStudent. Noisy Student (Xie et al., 2020b) is an iterative training method that leverages a teacher-student paradigm. The teacher model is first trained on annotated data. It is then asked to make predictions on the unlabeled data as pseudo-labels. Then, another student model with an equal or larger number of parameters is trained with a mixture of annotated and pseudo-labeled data. Note that pseudo labels, in numerical values, are directly used as the targeting labels. The trained student model will serve as a new teacher and re-label the unlabeled data again to yield a better prediction. In our implementation, dropout or dynamic model depth is introduced as noise to the model. All models θ are trained with standard cross-entropy loss, as shown in Equation 1. We set the dropout probability to 0.5, as it leads to the fastest convergence on our data. Only one iteration is completed in our experiment, as that's when the student model reaches its best result.

PseudoReasoner. PseudoReasoner (Fang et al., 2022) is another iterative semi-supervised learning framework that is proposed to tackle Commonsense Knowledge Base Population (CKBP) task (Fang et al., 2021a, 2023). It leverages a similar teacher-student paradigm and a novel filtering mechanism with the assistance of the student model. We replaced the generative teacher model with a DeBERTa-v3-large model due to the disastrous performance that GPT2 achieved on both verification tasks. Similar to CAT, two thresholds, $T^+ = 0.9$ and $T^- = 0.1$, are determined to assign pseudo-labels to unlabeled data based on the prediction of the teacher model. The rest steps remain the same as described in the original paper. Similar to NoisyStudent, only one iteration is carried out for PseudoReasoner as the student model converges to the best.

C.1.2 Settings

We use pretrained language models from the Huggingface Transformers[4] Library (Wolf et al., 2020) to build our framework. The learning rate for all models is set as 5e-6, and the batch size is 64. We use an AdamW (Loshchilov and Hutter, 2019) optimizer and evaluate the model every 25 steps. The max sequence length for the tokenizer is set to 25 and 35 for both discriminative tasks, respectively. Due to the imbalanced dataset, we evaluate the discriminative models with Area Under Curve (AUC) score (Bradley, 1997). Early stopping is used where the best checkpoint is selected when the largest validation AUC is achieved. All experiments are repeated three times using different random seeds, and the average performances and standard deviations are reported. In addition, we set the probability thresholds for both tasks to $T^+ = 0.9$ and $T^- = 0.1$ to determine the pseudo labels. The thresholds are roughly derived by observing the overall distribution and quality of data satisfying the respective threshold. For the bootstrapping method, we bootstrap $m = 9$ additional concepts for event conceptualization verification and $n = 2$ additional instances for abstract triple verification. Detailed ablation studies are provided in Section 5.3. As for the computational infrastructure, the models are trained and evaluated on four NVIDIA RTX3090 (24G) and four NVIDIA 1080Ti (12G) graphical cards. The number of parameters for every model is reported in Table 11.

C.1.3 Additional Experiment Results

The full experiment results for discriminative CSKB conceptualization tasks are reported in Table 11. All supervised learning baselines achieve comparable results as reported by He et al. (2022). Supervised CAT will be discussed later. The results by semi-supervised CAT are generally consistent with our findings as discussed in Section 5.1. To study the effect of different components and the training regime of CAT, we conduct more detailed ablation studies in Appendix C.1.4.

C.1.4 Ablation Study

In this section, we study the effects of different components in CAT and the training strategy of

[4] https://huggingface.co/docs/transformers

CAT. These studies indicate that our framework design and the proposed bootstrapping method play an important role in CSKB conceptualization and are more effective than leveraging unlabeled data with pseudo labels.

Framework Components. Our CAT framework consists of three critical components that make CAT different from traditional semi-supervised baselines. They are denoted as:

• Bootstrapping: Assist the training of student models by retrieving alternative conceptualizations and instantiations and bootstrapping them via natural language prompts. Dropping this component will train student models with the original textual prompts that are also used by the teacher models.

• CAT Cycle: Unite event and triple conceptualization tasks by assigning negative pseudo labels to abstract triples whose conceptualized head is predicted as wrong conceptualization. Dropping this component will separate the framework into two lines of training, which are training event conceptualization and triple conceptualization models separately.

• Pseudo-label refinement: Refine the pseudo labels with the latest student models and re-train the student models. Dropping this component will not update any pseudo label and will not re-train the student model.

Models	Event.	Triple.
CAT (BERT-base)	**87.4**	**76.3**
◇ w/o Bootstrapping	83.1	73.0
◇ w/o CAT Cycle	86.5	75.1
◇ w/o Pseudo-label Refinement	87.4	76.2
CAT (DeBERTa-v3-large)	**89.2**	**80.0**
◇ w/o Bootstrapping	84.0	77.7
◇ w/o CAT Cycle	88.1	79.0
◇ w/o Pseudo-label Refinement	89.1	79.7

Table 8: Ablation study on three components of CAT. Three components refer to the explanations above. The column **Event.** indicates test set AUC on the event conceptualization task, and the column **Triple.** indicates test set AUC on the triple conceptualization task.

We then conduct ablation studies regarding these three components with semi-supervised CAT to prove the effectiveness of our framework design and proposed bootstrapping method. Each component is removed separately, and the test set performances by student models are reported. The results are shown in Table 8. From the results, bootstrapping alternative conceptualization and instantiation leads to the largest performance gain. Bridging event conceptualization discrimination with triple conceptualization also causes slight improvements. However, refining the pseudo labels and re-train the student models have barely any effect. Thus, our bootstrapping method is the most important component within the entire CAT framework and can effectively assist in learning conceptual knowledge.

Supervised CAT. We further study training CAT in a supervised learning setting to examine the role of unlabeled data. In supervised CAT, no teacher models are trained to provide pseudo labels. The alternative conceptualizations and instantiations are retrieved directly from the annotated event conceptualization data and bootstrapped later. Two student models are trained on the bootstrapped data only and evaluated on the same testing set, and the results are reported in Table 11. Compared with supervised learning baselines, supervised CAT can achieve a comparable result on the event conceptualization task. This may be due to the fact that the diversity of concepts drops without considering unlabeled conceptualizations. Improvements in the triple conceptualization task are more significant, and the results are comparable with semi-supervised CAT. This indicates that our framework design and bootstrapping method are successful in discriminating high-quality abstract commonsense knowledge, and leveraging a semi-supervised learning paradigm benefits more in event conceptualization discrimination.

C.2 Application and Evaluation of CAT

C.2.1 Settings

Pretrained GPT2 models from the Huggingface Transformers Library and training codes[5] by Hwang et al. (2021) are used as our code base. The learning rate for all experiments is set to 1e-5, and the batch size is fixed to 64. We use an Adam (Kingma and Ba, 2015) optimizer and evaluate the model every 20 steps. The input and output lengths for GPT2 models are fixed at 45 and 55 for the two application and evaluation tasks, respectively. Such length settings can cover all annotated conceptualizations and triples. For both generative experiments, we evaluate the generations with BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), ROUGE-L (Lin, 2004), and

[5]https://github.com/allenai/comet-atomic-2020

CIDEr (Vedantam et al., 2015) scores. However, since an abstract concept usually contains one or two tokens, we only report BLEU1 and BLEU2 scores for the generative event conceptualization task. Early stopping is also applied where the best checkpoint is selected when the minimum autoregressive LM loss is achieved. In addition, we notice that the number of triples from the ATOMIC subset is much smaller than abstract triples for the commonsense inference modeling task. Thus, we upsample the ATOMIC subset by a ratio of 1:2 across all experiments to guarantee a consistent and balanced number of training data. For generative event conceptualization, the training data is simply a mixture of annotated and pseudo-labeled event conceptualizations without any balancing measure. All the models are trained and evaluated on four NVIDIA RTX A6000 graphical cards with 48G memory. The number of parameters is close to the number of parameters in GPT2-XL, which is reported in Table 11.

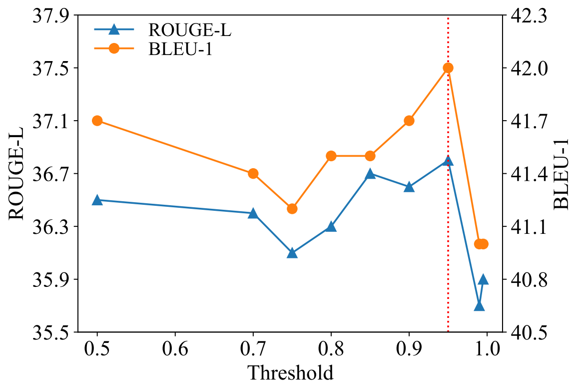

Figure 5: Performance (%) curve by COMET (GPT2-XL) on commonsense inference generation task with different thresholds for determining positive pseudo labels. Performance with the best threshold of 0.95 is marked as the red dotted line.

C.2.2 Annotation Settings

When evaluating the event conceptualization generator, expert annotations are conducted to evaluate concepts that are not presented in the training set. Crowdsourced platforms such as Amazon Mechanical Turk are not used since experts understand conceptualization better and are more reliable for evaluation. Subsequently, the authors of this paper are invited to serve as expert annotators. They are experienced in NLP research and clearly understand the paper's scope. The annotation guideline is carefully designed. Each question presents the original head event with the instance highlighted and the corresponding conceptualization candidate to be annotated. There are also several positive and negative conceptualizations attached as examples. The authors are well-informed about the instruction and the intended use of their annotations in this paper. And they all agreed to annotate as part of their contributions. Moreover, in order to ensure that the expert will not deliberately raise the plausible rate of a certain set of annotation candidates, we randomly shuffle all the data and invite one more expert to cross-validate the annotations. These measures can ensure that the annotation process is free of ethical concerns and justifiable.

C.2.3 Additional Experiment Results

We conduct a more comprehensive study on the commonsense inference generation task by experimenting with the effect of threshold tuning when filtering abstract commonsense knowledge. Multiple thresholds ranging from 0.5 to 0.995 are experimented with to derive abstract commonsense knowledge of different qualities. COMET (GPT2-XL) generators are fine-tuned on the ATOMIC subset, augmented by a mixture of annotated and pseudo-labeled abstract triples. The performance curve according to the threshold is plotted in Figure 5. Full version results with all metrics are reported in Table 19. It can be observed that gradually increasing the threshold from 0.75 will lead to better performance, which may be due to the improvement in data quality. However, increasing the threshold over 0.95 will cause a performance drop. One possible reason is the amount of pseudo-labeled triples significantly drops with a relatively high threshold, and COMET fails to learn well from annotated triples only. Using the CAT framework to pseudo-label unlabeled abstract triples leads to better performance than leveraging a RoBERTa-large supervised discriminator to assign pseudo-labels, which also validates the reliability of the triple conceptualization discriminator in CAT. Also, it is noticeable that training COMET with triples based on our constructed ATOMIC subset is much worse than training with the full ATOMIC dataset. This indicates that exposing the model with substantial factual commonsense knowledge is still important, and only equipping the model with abstract commonsense knowledge is not enough for commonsense inference modeling.

D Computational Cost Analysis

In this section, we compare the number of training data used for both CSKB conceptualization tasks to compare the computational cost across different frameworks and methodologies empirically. Both annotated and pseudo-labeled data are counted. The comparison result is presented in Table 9. All semi-supervised learning methods leverage a significant amount of unlabeled data due to the great scarcity of annotations. With threshold filterings, PseudoReasoner (Fang et al., 2022) and our CAT framework can abandon more than half of pseudo examples with poor quality. Even though our CAT framework can still outperform PseudoReasoner and achieve the best performance among all methods. Additionally, there is no notable increase in the number of model parameters as CAT also applies a teacher-student paradigm that is similar to Noisy-Student and PseudoReasoner. Even compared with the supervised baselines, CAT only doubles the parameters used. In conclusion, with comparable training data and parameters against other baselines, CAT can achieve much better results and state-of-the-art performances.

Method	Event.	Triple.	Total
Supervised Baselines	107,384	65,386	172,770
UDA	412,367	4,916,658	5,329,025
Noisy-Student	412,367	4,916,658	5,329,025
PseudoReasoner	316,601	1,727,865	2,044,466
CAT	317,507	1,595,411	1,912,918

Table 9: Comparison between the number of training data for discriminative event conceptualization (Event.) and triple conceptualization (Triple.) tasks.

E Case Studies

This section contains case studies of the four tasks we studied in this paper, including CSKB conceptualization tasks and applications of CAT. Throughout these cases, we would like to offer a clearer view of the data, discuss the challenges of the conceptualization task, and provide brief error analyses.

E.1 CSKB Conceptualization

Event Conceptualization. For discriminative event conceptualization, the case study is shown in Table 15. From these cases, it can be observed that several instances i can be identified within one head event h_o, and each of them can be conceptualized in multiple ways. Formally, assume we are conceptualizing m events, each with n instances. And each instance i concerned can be conceptualized as p concepts. Each concept takes the majority vote of q annotators to verify. Subsequently, the number of annotations needed is $O(mnpq)$, which grows significantly if we conceptualize a commonsense knowledge base at scale. Thus, it is extremely infeasible for practitioners to annotate all of the conceptualizations for verification, which also highlights the importance of a reliable discriminative conceptualization model as CAT acquired. Semi-supervised learning is also an ideal training strategy, as there is a considerable amount of unlabeled data.

Analyzing the errors made by our discriminator, we observe that models frequently make errors when the instance contains the word “PersonX,” which could be caused by the reporting bias (Gordon and Durme, 2013), as “PersonX” is seldom used in normal natural language texts. Replacing the subjects with commonly used names such as “Alex, Bob” may alleviate such a problem. Additionally, models make errors on some rarely seen concepts, such as “organ,” “cognitive ability,” and “side effect.” Their absence from training data can partially cause this, as CSKB, like ATOMIC, may not cover many instances under those rarely used concepts.

Triple Conceptualization. For triple conceptualization discrimination, case studies are shown in Table 17. Similar to the analysis above, consider m events with n instances, each instance with p concepts. Assume that every ATOMIC head event has t relation and tail tuples as its counterpart, and q votes are required from annotators. The total number of annotations is $O(mnptq)$ for verifying all abstract commonsense triples, which is also huge compared with the total number of original commonsense triples.

The errors are mainly due to the loss of contextualization within the original head events, as conceptualized head events with too high abstractness are likely to omit salient properties. For example, conceptualizing “watching a scary movie” as “watching movie” will lose the property “scary,” which further leads to a wrong abstract commonsense knowledge if the tail is “feel scared.” This also highlights the importance of verifying the plausibility of abstract commonsense knowledge that heavily relies on both the contextualization brought by r, t and the conceptualization of the head event. Meanwhile, we observe that the models tend to make a neutral decision (plausibility score close to 0.5) when encountering the situation of conceptualizing an entire event as a concept with high-level abstractness. Indeed, they are more difficult abstract commonsense knowledge for machines to learn, as a higher level of abstractness leads to more possible instantiations and commonsense inferences.

E.2 Appliaction of CAT

Generative Event Conceptualization. The examples are shown in Table 16. Generated conceptualizations are generally plausible, given the head event as the context. Specifically, we observe that neural generators are more sensitive to the instance and its context, as heuristic matching may conceptualize "sleeping at night" and "having trouble sleeping at night" as "sleeping". In contrast, neural generators can distinguish these two instances clearly by conceptualizing them as "sleep" and "sleep disorder". One potential weakness of neural generators is that the generated conceptualizations lack diversity and novelty (Du et al., 2019; Wang et al., 2021), as they tend to be semantically close to the targeting conceptualizations in the training samples. Nevertheless, it still offers a reliable and simplified approach to performing contextualized conceptualization without tedious matching and human annotations. Such results also validate the reliability of our discriminative event conceptualization model, as the pseudo-labeled conceptualizations tend to be of high quality.

Commonsense Inference Modeling (COMET). Generations from COMET that are only trained on the ATOMIC subset, possibly augmented by abstract commonsense triples, are compared in Table 18. From these generations, we can observe that the abstract commonsense knowledge-aided COMET generator can generate tail events that are most plausible and generalizable compared with the one only trained on ATOMIC. It generally supports our hypothesis that abstract commonsense knowledge may implicitly help model situational commonsense inference, even without the instantiation step. In addition, this also validates that our automatically derived abstract knowledge is reliable and helpful, which also proves the reliability of our triple conceptualization discriminator.

E.3 Conceptualization by Large Language Models

With the recent advances in Large Language Models (LLMs), such as GPT3.5 (Brown et al., 2020; Ouyang et al., 2022) and ChatGPT (OpenAI, 2022), on various NLP tasks (Qin et al., 2023; Bian et al., 2023; Chan et al., 2023; Amin et al., 2023), we also aim to explore ChatGPT's conceptualization ability through case studies. To do so, we investigate ChatGPT's performance on three conceptualization tasks: discriminative event conceptualization, discriminative triple conceptualization, and generative event conceptualization, all of which are defined in Section 3. We randomly sample data entries from AbstractATOMIC and prompt ChatGPT with natural language commands to perform various tasks. The prompts used for performing these tasks are listed in Table 10. Specifically, we use OpenAI's API[6] to prompt ChatGPT and retrieve its generations.

The case studies for three tasks are presented in Table 12, Table 13, and Table 14, respectively. These demonstrate ChatGPT's strong conceptualization abilities in both discriminative and generative manners. While ChatGPT can accurately determine most event conceptualizations and abstract commonsense knowledge, it still makes some mistakes. This highlights the value of training a performant discriminator through CAT, as it can effectively detect incorrect conceptualizations and implausible abstract commonsense knowledge. Additionally, ChatGPT tends to conceptualize instances using synonyms (Hagiwara et al., 2006) and hypernyms (Yu et al., 2020) and paraphrased or explained terms rather than higher-level concepts. This underscores the importance of our event conceptualization generator, which can generate precise, concise event conceptualizations. In conclusion, our work holds significant value in the realm of commonsense reasoning through conceptualization, particularly in light of the rise of large language models.

[6]The code for the model is `gpt-3.5-turbo`, and the date of access is May 2023.

Task	Prompt
Discriminative Event Conceptualization	`Given the event` <*event*>`, can the` <*instance*> `be conceptualized as` <*concept*>`? Only answer yes or no without any other words. You are forced to make a decision.`
Discriminative Triple Conceptualization	`Given a commonsense knowledge triple,` <*head, relation, tail*>`, is this knowledge plausible or not? Only answer yes or no without any other word. You are forced to make a decision.`
Generative Event Conceptualization	`Given the event` <*event*>`, what are possible conceptualizations of` <*instance*>`? Only list out five short conceptualizations, and do not provide explanations.`

Table 10: Natural language prompts used to instruct ChatGPT to perform specific tasks. Words in italics and enclosed by brackets indicate inputs replaced by sampled data entries. Restrictive commands are appended at the end to ensure ChatGPT executes the task as intended.

Framework	Backbone PTLM / Method	Event Conceptualization		Triple Conceptualization	
		Validation	Testing	Validation	Testing
Supervised Learning	BERT-base *110M*	82.4$_{\pm0.05}$	82.5$_{\pm0.31}$	71.2$_{\pm0.58}$	72.6$_{\pm0.71}$
	BERT-large *340M*	82.8$_{\pm0.48}$	83.1$_{\pm0.80}$	72.4$_{\pm0.01}$	73.7$_{\pm0.00}$
	BART-base *139M*	83.8$_{\pm0.28}$	84.4$_{\pm0.32}$	72.0$_{\pm0.09}$	72.6$_{\pm0.15}$
	BART-large *406M*	85.0$_{\pm0.13}$	85.2$_{\pm0.22}$	74.5$_{\pm0.13}$	76.2$_{\pm0.19}$
	RoBERTa-base *110M*	84.1$_{\pm0.04}$	84.5$_{\pm0.19}$	72.2$_{\pm0.00}$	74.1$_{\pm0.00}$
	RoBERTa-large *340M*	85.2$_{\pm0.24}$	85.5$_{\pm0.02}$	75.3$_{\pm0.00}$	76.9$_{\pm0.01}$
	DeBERTa-v3-base *214M*	85.1$_{\pm0.08}$	85.8$_{\pm0.07}$	73.9$_{\pm0.10}$	75.9$_{\pm0.04}$
	DeBERTa-v3-large *435M*	85.8$_{\pm0.05}$	86.2$_{\pm0.15}$	76.9$_{\pm0.03}$	78.0$_{\pm0.02}$
	ELECTRA-base *110M*	85.4$_{\pm0.05}$	85.8$_{\pm0.02}$	74.3$_{\pm0.27}$	76.2$_{\pm0.12}$
	ELECTRA-large *340M*	84.7$_{\pm0.47}$	85.3$_{\pm0.38}$	75.6$_{\pm0.01}$	77.9$_{\pm0.06}$
	GPT2-base *117M*	60.0$_{\pm0.06}$	59.1$_{\pm0.14}$	52.8$_{\pm0.14}$	55.9$_{\pm0.11}$
	GPT2-medium *345M*	61.2$_{\pm0.11}$	60.3$_{\pm0.08}$	54.6$_{\pm0.17}$	57.4$_{\pm0.09}$
	GPT2-large *774M*	64.1$_{\pm0.05}$	62.7$_{\pm0.08}$	60.5$_{\pm0.11}$	59.8$_{\pm0.06}$
	GPT2-XL *1558M*	64.2$_{\pm0.19}$	63.6$_{\pm0.22}$	62.2$_{\pm0.08}$	61.5$_{\pm0.10}$
Semi-Supervised Learning	UDA (TF-IDF)	83.6$_{\pm0.29}$	83.6$_{\pm0.24}$	75.8$_{\pm1.26}$	76.8$_{\pm1.34}$
	UDA (back-trans.)	83.4$_{\pm0.27}$	83.6$_{\pm0.24}$	75.8$_{\pm1.25}$	76.8$_{\pm1.34}$
	Noisy-Student	86.4$_{\pm0.05}$	86.5$_{\pm0.09}$	75.4$_{\pm0.64}$	76.7$_{\pm0.59}$
	PseudoReasoner (BERT-base)	83.3$_{\pm0.11}$	84.0$_{\pm0.24}$	73.0$_{\pm0.14}$	74.1$_{\pm0.33}$
	PseudoReasoner (RoBERTa-large)	86.6$_{\pm0.25}$	86.7$_{\pm0.33}$	76.3$_{\pm0.12}$	77.2$_{\pm0.21}$
CAT ***(Supervised)***	BERT-base *110M*	83.9$_{\pm0.42}$	84.5$_{\pm0.43}$	73.4$_{\pm0.32}$	73.3$_{\pm0.23}$
	BERT-large *340M*	82.8$_{\pm0.48}$	83.1$_{\pm0.80}$	72.4$_{\pm0.01}$	73.7$_{\pm0.00}$
	BART-base *139M*	84.9$_{\pm0.05}$	85.4$_{\pm0.08}$	75.2$_{\pm0.06}$	76.9$_{\pm0.21}$
	BART-large *406M*	86.2$_{\pm0.05}$	86.0$_{\pm0.06}$	76.8$_{\pm0.21}$	78.7$_{\pm0.31}$
	RoBERTa-base *110M*	85.5$_{\pm0.06}$	86.0$_{\pm0.06}$	76.6$_{\pm0.12}$	77.2$_{\pm0.18}$
	RoBERTa-large *340M*	86.2$_{\pm0.31}$	86.2$_{\pm0.31}$	77.7$_{\pm0.19}$	78.5$_{\pm0.28}$
	DeBERTa-v3-base *214M*	85.8$_{\pm0.15}$	86.2$_{\pm0.07}$	76.8$_{\pm0.28}$	79.0$_{\pm0.20}$
	DeBERTa-v3-large *435M*	86.3$_{\pm0.11}$	86.7$_{\pm0.08}$	78.4$_{\pm0.20}$	79.5$_{\pm0.18}$
	ELECTRA-base *110M*	85.5$_{\pm0.12}$	85.7$_{\pm0.08}$	76.7$_{\pm0.05}$	77.3$_{\pm0.16}$
	ELECTRA-large *340M*	86.2$_{\pm0.66}$	86.0$_{\pm0.62}$	77.8$_{\pm0.11}$	78.5$_{\pm0.09}$
CAT ***(Semi-Supervised)***	BERT-base *110M*	87.1$_{\pm0.06}$	87.4$_{\pm0.11}$	74.3$_{\pm0.26}$	76.3$_{\pm0.38}$
	BERT-large *340M*	87.7$_{\pm0.16}$	88.0$_{\pm0.19}$	75.8$_{\pm0.23}$	77.8$_{\pm0.36}$
	BART-base *139M*	88.2$_{\pm0.09}$	88.2$_{\pm0.09}$	75.7$_{\pm0.09}$	78.0$_{\pm0.14}$
	BART-large *406M*	88.6$_{\pm0.07}$	88.7$_{\pm0.10}$	77.2$_{\pm0.12}$	79.0$_{\pm0.14}$
	RoBERTa-base *110M*	88.4$_{\pm0.12}$	88.3$_{\pm0.08}$	76.9$_{\pm0.16}$	78.0$_{\pm0.19}$
	RoBERTa-large *340M*	89.0$_{\pm0.15}$	88.8$_{\pm0.20}$	78.2$_{\pm0.08}$	79.4$_{\pm0.14}$
	DeBERTa-v3-base *214M*	88.8$_{\pm0.12}$	88.9$_{\pm0.08}$	77.5$_{\pm0.10}$	79.9$_{\pm0.07}$
	DeBERTa-v3-large *435M*	**89.1$_{\pm0.05}$**	**89.2$_{\pm0.14}$**	**78.7$_{\pm0.16}$**	**80.0$_{\pm0.33}$**
	ELECTRA-base *110M*	88.7$_{\pm0.10}$	88.9$_{\pm0.10}$	74.9$_{\pm0.15}$	75.5$_{\pm0.40}$
	ELECTRA-large *340M*	88.6$_{\pm0.77}$	88.5$_{\pm0.70}$	74.9$_{\pm0.15}$	75.5$_{\pm0.40}$

Table 11: Full experiment results (%) by our CAT framework on the discriminative event conceptualization and triple conceptualization tasks. We report the average AUC score and standard deviation across experiments with three random seeds. The best performances within each framework are underlined, and the best among all models are bold-faced. All supervised baselines are comparable with experiment results by He et al. (2022).

Head Event	Instance	Concept	Label	Pred.
PersonX accepts the invitation	the invitation	personal communication	✓	✓
	the invitation	party idea	×	✓
	the invitation	friendly approach	✓	✓
	the invitation	item	×	✓
	PersonX accepts the invitation	acceptance	✓	✓
	PersonX accepts the invitation	approach	×	×
	PersonX accepts the invitation	psychological treatment	×	×
	PersonX accepts the invitation	personal communication	✓	✓
PersonX makes oatmeal for breakfast	oatmeal	ingredient	×	✓
	oatmeal	cereal	✓	✓
	oatmeal	grain food	✓	✓
	breakfast	service	×	×
	breakfast	meal	✓	✓
	PersonX makes oatmeal for breakfast	hands-on activity	✓	✓
	PersonX makes oatmeal for breakfast	extended school activity	×	✓
	PersonX makes oatmeal for breakfast	cooking	✓	✓

Table 12: Case study of ChatGPT's discriminative event conceptualizations. Label refers to annotation result and Pred. stands for prediction by ChatGPT.

Conceptualized Head Event	Relation	Tail Event	Label	Pred.
medical check	xEffect	to be brave	✓	×
	xWant	take medicine	✓	✓
	xWant	leave the hotel	×	✓
	xWant	to drive home	×	✓
	xAttr	cautious	✓	✓
	xWant	go to rest	✓	✓
	xAttr	diseased	✓	×
	xNeed	get injured	✓	×
watching movie	xEffect	laugh	✓	×
	xWant	to be entertained	✓	✓
	xNeed	to go to video store	✓	×
	xIntent	entertain themselves	✓	✓
	xWant	to put movie in DVD player	✓	✓
	xAttr	satisfied	✓	×
	xReact	scared	✓	✓
	xNeed	rent it on Netflix	✓	✓

Table 13: Case study of ChatGPT's discriminative triple conceptualizations. Underlined words in the head event refer to conceptualizations. Label refers to annotation result and Pred. stands for prediction by ChatGPT.

Event	Target Conceptualizations	CAT's Generations	ChatGPT's Generations
PersonX is having trouble sleeping at night	time, night, nonwork time, shift	night, evening, time, late, darknight	Dark sleep time, nocturnal period, resting hours, nighttime sleeplessness, bedtime wakefulness
PersonX is having trouble sleeping at night	relaxation, sleeping, resting	rest, sleep, resting, daily routine, sleeping	Nightly slumber attempt, dark rest struggle, sleeplessness after dark, nocturnal insomnia, bedtime wakefulness
PersonX is having trouble sleeping at night	sleeping, disorder, sleep problem, trouble, insomnia, sleep disorder	sleep disorder, problem, sleep disturbance, difficulty, trouble sleeping	Nighttime sleep difficulty, dark restlessness problem, nocturnal insomnia struggle, bedtime wakefulness issue, sleeplessness after dark challenge

Table 14: Case study of ChatGPT's generative event conceptualizations. The instance candidate in each event is underlined. Target conceptualizations are positive conceptualizations extracted from AbstractATOMIC, including the annotated conceptualizations and ones that are positively pseudo-labeled by our framework.

Head Event	Instance	Concept	Label	Pred.
PersonX is having trouble sleeping at night	night	nonwork time	✓	✓
	night	night	✓	✓
	sleeping at night	lifestyle factor	✓	×
	sleeping at night	basic need	✓	✓
	trouble sleeping at night	board game	×	×
	trouble sleeping at night	problem	✓	✓
	PersonX is having trouble sleeping at night	variable	×	×
	PersonX is having trouble sleeping at night	personal characteristic	✓	✓
PersonX is nervous about making friends	friends	person	✓	✓
	friends	support person	✓	✓
	making friends	relationship	✓	✓
	making friends	social activity	✓	✓
	nervous about making friends	organ	×	✓
	nervous about making friends	side effect	×	✓
	PersonX is nervous about making friends	emotion	✓	✓
	PersonX is nervous about making friends	nervous disorder	✓	✓
PersonX wants to learn how to play the piano	the piano	instrument	✓	✓
	the piano	western instrument	✓	✓
	how to play the piano	musical activity	✓	✓
	how to play the piano	play	✓	×
	to learn how to play the piano	button	×	×
	to learn how to play the piano	learning activity	✓	✓
	PersonX wants to learn how to play the piano	cultural event	×	×
	PersonX wants to learn how to play the piano	cognitive ability	✓	×
PersonX puts PersonX's pants on PersonX's leg at a time	PersonX's pants	pant	✓	×
	PersonX's pants	clothing	✓	✓
	PersonX's leg	leg	✓	×
	PersonX's leg	limb	✓	×
	a time	resource	×	×
	a time	time	✓	✓
	PersonX puts PersonX's pants on PersonX's leg	dressing	✓	✓
	PersonX puts PersonX's pants on PersonX's leg	action	×	×

Table 15: Case study of CAT's discriminative event conceptualizations. A head event can be conceptualized in multiple ways, as shown in the table. Label refers to annotation result and Pred. stands for prediction by our framework.

Event	Target Conceptualizations	Generated Conceptualizations
PersonX is having trouble sleeping at night	time, night, nonwork time, shift	night, evening, time, late, darknight
PersonX is having trouble sleeping at night	relaxation, sleeping, resting	rest, sleep, resting, daily routine, sleeping
PersonX is having trouble sleeping at night	sleeping, disorder, sleep problem, trouble, insomnia, sleep disorder	sleep disorder, problem, sleep disturbance, difficulty, trouble sleeping
PersonX gets great grades in school	accomplishment, result, grades, good performance, achievement	achievement, grades, good grade, academic excellence, grade
PersonX asks what was wrong	problems, concern, seeking information, questioning, query, communication	query, question, asking, communication, inquiry
PersonX needs new shoes	necessity, product, personal item, item, clothing, shoes	requirement, item, need, necessity, needs
PersonX is failing math	negative experience, negative issue, problem, poor performance	difficulty, poor performance, problem, academic failure, math problem

Table 16: Case study of CAT's generative event conceptualizations. The instance candidate in each event is underlined. Target conceptualizations are positive conceptualizations extracted from AbstractATOMIC, including the annotated conceptualizations and ones that are positively pseudo-labeled by our framework.

Conceptualized Head Event	Relation	Tail Event	Label	Pred.
PersonX gets nailcare service	xAttr	rich	✓	✓
	xAttr	skillful	×	✓
	xIntent	look pretty	✓	✓
	xNeed	book an appointment	✓	✓
	xEffect	show off	✓	✓
	xReact	excited	✓	✓
	oWant	to tell her they like them	✓	✓
	xWant	to go home	✓	✓
watching movie	xEffect	laugh	✓	✓
	xWant	to be entertained	✓	✓
	xNeed	to go to video store	✓	×
	xIntent	entertain themselves	✓	✓
	xWant	to put movie in DVD player	✓	✓
	xAttr	satisfied	✓	✓
	xReact	scared	✓	×
	xNeed	rent it on Netflix	✓	✓
medical check	xEffect	to be brave	✓	✓
	xWant	take medicine	✓	✓
	xWant	leave the hotel	×	✓
	xWant	to drive home	×	✓
	xAttr	cautious	✓	✓
	xWant	go to rest	✓	✓
	xAttr	diseased	✓	✓
	xNeed	get injured	✓	✓

Table 17: Case study of CAT's discriminative triple conceptualizations. The abstract concept within each conceptualized head event is underlined. Label refers to annotation result and Pred. stands for prediction by our framework.

Head	Relation	Source	Tail
PersonX washes PersonY's car	oWant	ATOMIC	to tip PersonX
		$\text{COMET}_{\text{ATOMIC}}$	to wash their car
		$\text{COMET}_{\text{CAT}}$	to thank PersonX
PersonX meets PersonX's standards	xNeed	ATOMIC	to practice
		$\text{COMET}_{\text{ATOMIC}}$	to study
		$\text{COMET}_{\text{CAT}}$	to practice hard
PersonX stretches out PersonX's hand	xWant	ATOMIC	to give PersonY something
		$\text{COMET}_{\text{ATOMIC}}$	to touch
		$\text{COMET}_{\text{CAT}}$	to grab something for PersonY
PersonX learns how to bake a cake	xAttr	ATOMIC	interested
		$\text{COMET}_{\text{ATOMIC}}$	curious
		$\text{COMET}_{\text{CAT}}$	skilled
PersonX fails PersonX's class	xWant	ATOMIC	to retake the class
		$\text{COMET}_{\text{ATOMIC}}$	to study hard
		$\text{COMET}_{\text{CAT}}$	to try again in the class
PersonX buys dog food	xEffect	ATOMIC	X gets receipt
		$\text{COMET}_{\text{ATOMIC}}$	loses weight
		$\text{COMET}_{\text{CAT}}$	gets a receipt
PersonX hits by lightning	xEffect	ATOMIC	has hair burned
		$\text{COMET}_{\text{ATOMIC}}$	gets electrocuted
		$\text{COMET}_{\text{CAT}}$	screams in pain
PersonX forgets my wallet	xEffect	ATOMIC	is chastised
		$\text{COMET}_{\text{ATOMIC}}$	gets robbed
		$\text{COMET}_{\text{CAT}}$	thinks about it
PersonX realizes something	xWant	ATOMIC	make a plan
		$\text{COMET}_{\text{ATOMIC}}$	to solve the problem
		$\text{COMET}_{\text{CAT}}$	to do something about it

Table 18: Case study of commonsense inference generation (COMET). Examples are selected from the original ATOMIC testing set. ATOMIC refers to the target tail in the original ATOMIC. $\text{COMET}_{\text{ATOMIC}}$ and $\text{COMET}_{\text{CAT}}$ stand for generations by COMET trained on an ATOMIC subset or aided with abstract knowledge derived by CAT.

Training Data	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Zero-Shot	5.42	4.89	1.84	1.51	0.65	0.52	0.26	0.21	6.50	5.70	6.40	5.90	1.60	1.20
ATOMIC (subset)	38.1	38.1	25.4	25.7	18.7	18.8	15.5	15.7	14.9	14.9	33.0	33.2	27.6	27.8
$+D^l_t$	38.1	38.5	24.8	25.5	17.8	18.4	14.7	15.2	15.3	15.6	33.1	33.7	26.8	27.3
+Finetune	38.6	39.0	25.8	26.6	18.9	19.7	15.7	16.4	15.1	15.4	33.6	34.4	28.8	30.0
$+D^u_{\text{Abs.ATM.}}$	40.0	40.3	27.1	27.8	20.0	20.8	16.5	17.5	16.1	16.3	35.3	35.7	31.6	31.7
+Finetune	40.1	40.5	27.1	27.8	20.1	20.8	16.7	17.4	16.2	16.4	35.4	35.9	31.8	31.7
$+D^l_t + D^u_{\text{Abs.ATM.}}$	40.2	40.6	26.2	27.4	19.0	20.4	15.1	16.8	16.3	16.5	35.0	35.4	31.0	31.3
+Finetune	40.0	40.4	26.0	26.9	18.7	19.7	15.0	16.1	16.3	16.4	35.0	35.4	30.3	30.7
$+D^u_{0.995}$	39.7	39.8	26.5	26.8	19.5	19.8	15.6	16.1	15.8	15.8	35.0	34.9	30.8	30.7
+Finetune	41.0	41.0	27.1	27.5	20.0	20.2	16.1	16.3	16.7	16.6	36.0	35.9	31.9	31.7
$+D^u_{0.99}$	39.5	39.9	26.1	27.0	19.3	20.0	15.9	16.6	15.7	15.9	34.7	34.8	30.6	30.8
+Finetune	40.8	41.0	27.0	27.6	20.0	20.5	16.2	16.9	16.7	16.6	35.8	35.7	31.9	31.6
$+D^u_{0.95}$	41.2	41.9	28.1	29.0	20.7	21.5	16.5	17.8	16.6	16.9	35.9	36.5	33.4	33.7
+Finetune	41.1	42.0	28.0	29.0	20.4	21.5	16.4	17.6	16.6	17.0	36.0	36.8	33.2	33.8
$+D^u_{0.90}$	41.6	41.6	28.1	28.5	20.9	21.5	17.1	17.7	16.9	16.8	36.7	36.4	33.4	33.1
+Finetune	41.8	41.7	28.3	28.5	21.0	21.4	17.0	17.5	17.0	17.0	36.7	36.6	33.4	33.1
$+D^u_{0.85}$	41.3	41.4	27.8	28.1	20.7	21.1	16.8	17.6	16.7	16.8	36.3	36.6	32.6	32.9
+Finetune	41.5	41.5	27.9	28.2	20.6	21.1	16.8	17.5	16.8	16.9	36.3	36.7	32.6	33.0
$+D^u_{0.80}$	41.6	41.6	27.3	28.0	20.1	20.7	16.3	17.0	17.0	16.9	36.6	36.4	33.0	32.6
+Finetune	41.6	41.5	27.5	27.9	20.2	20.6	16.3	16.8	17.0	16.9	36.6	36.3	33.0	32.3
$+D^u_{0.75}$	40.6	40.8	27.1	28.0	19.9	20.9	16.2	17.2	16.4	16.6	35.5	35.7	31.6	32.1
+Finetune	40.9	41.2	27.2	28.1	19.9	21.0	16.2	17.0	16.6	16.9	35.7	36.1	31.8	32.7
$+D^u_{0.70}$	40.6	40.9	27.1	27.8	19.9	20.7	16.6	17.2	16.4	16.6	35.6	36.1	31.6	32.4
+Finetune	41.4	41.4	27.5	28.1	20.1	21.0	16.4	17.4	16.9	16.9	36.2	36.4	32.5	33.0
$+D^u_{0.50}$	41.1	41.5	27.3	28.2	20.4	21.2	16.7	17.6	16.7	16.7	35.8	36.1	32.4	32.8
+Finetune	41.5	41.7	27.7	28.5	20.7	21.4	17.0	17.8	16.9	16.9	36.3	36.5	32.7	33.1
$+D^l_t + D^u_{0.995}$	39.4	39.3	26.1	26.4	19.2	19.5	15.5	15.8	15.7	15.5	33.9	33.8	29.8	29.2
+Finetune	39.7	40.0	26.7	27.5	19.5	20.3	15.8	16.6	15.7	15.7	34.7	34.9	30.6	30.9
$+D^l_t + D^u_{0.99}$	39.4	39.7	25.7	26.5	18.6	19.5	15.2	16.5	15.8	15.9	34.6	35.0	29.7	30.2
+Finetune	39.7	40.4	26.6	27.6	19.6	20.5	16.0	16.8	15.7	16.1	34.2	35.0	30.5	31.1
$+D^l_t + D^u_{0.95}$	39.9	40.5	26.2	27.4	19.3	20.6	16.0	17.4	16.0	16.2	35.0	35.4	30.8	31.3
+Finetune	40.4	41.0	26.6	27.6	19.5	20.7	16.1	17.1	16.2	16.5	35.4	35.8	31.3	31.5
$+D^l_t + D^u_{0.90}$	39.4	39.7	26.1	27.0	18.9	19.9	15.3	16.4	15.6	15.8	34.5	35.0	29.6	30.2
+Finetune	40.4	40.4	26.2	26.9	19.1	19.6	15.2	15.8	16.3	16.4	35.5	35.7	30.5	30.7
$+D^l_t + D^u_{0.85}$	39.8	40.0	26.3	26.9	19.3	19.8	15.8	16.1	16.0	16.2	34.8	35.2	30.5	30.6
+Finetune	39.9	40.0	26.2	26.7	19.3	19.5	15.8	15.8	16.1	16.3	34.9	35.5	30.4	30.7
$+D^l_t + D^u_{0.80}$	39.9	40.4	26.4	27.6	19.2	20.5	15.4	16.8	16.2	16.3	34.9	35.3	30.3	31.3
+Finetune	39.9	40.4	26.2	27.5	18.9	20.3	15.2	16.7	16.2	16.5	35.0	35.6	30.2	31.3
$+D^l_t + D^u_{0.75}$	39.7	39.8	25.9	26.6	18.9	19.4	15.3	15.8	15.6	15.7	34.6	34.9	29.7	30.1
+Finetune	39.8	39.9	25.9	26.7	18.8	19.5	15.3	15.9	15.7	15.9	34.7	35.1	29.6	30.3
$+D^l_t + D^u_{0.70}$	40.2	40.5	26.4	27.2	19.4	20.1	15.8	16.4	16.4	16.5	35.2	35.5	30.8	31.0
+Finetune	40.3	40.6	26.4	27.1	19.4	19.9	15.9	16.0	16.5	16.6	35.2	35.7	30.5	30.9
$+D^l_t + D^u_{0.50}$	39.3	39.8	26.2	27.5	18.9	20.3	15.2	16.7	15.7	16.0	33.9	34.4	29.4	30.6
+Finetune	39.5	40.1	26.3	27.6	19.0	20.5	15.4	17.1	15.8	16.2	34.2	34.9	29.3	30.8
ATOMIC (full)	42.7	42.9	29.6	30.0	22.0	22.5	18.6	18.7	29.1	29.7	51.1	52.7	74.5	75.4

Table 19: Full experiment results (%) by GPT2 (XL) on commonsense inference generation (COMET) task. We evaluate the models on the original ATOMIC dev and test sets. D^l_t stands for annotated abstract commonsense triples, and D^u stands for unlabeled triples pseudo-labeled by our CAT framework. The underfoot value is the threshold for selecting plausible pseudo labels. Fine-tune refers to fine-tuning back on the training set of our constructed ATOMIC subset. Rows with the best performance, which are reported in the paper, are colored in gray. We also report performances by COMET trained on the complete ATOMIC training set in the bottom row.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, in the limitation section on Page 10.

☑ A2. Did you discuss any potential risks of your work?
Yes, in the ethics statement section on Page 10.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, in both sections on the first page.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Yes, as introduced in Section 3, Problem Definition, and Section 5, Experiments. There is also an additional explanation in Appendix A, Dataset Description.

☑ B1. Did you cite the creators of artifacts you used?
Yes, all datasets are properly cited throughout the paper.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Yes, in the ethics statement section on Page 10, all datasets are shared via open-access licenses.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes, in the ethics statement section on Page 10, our use of existing artifacts is consistent with their intended use for research purposes.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Yes, in the ethics statement section on Page 10, the primary dataset is desensitized and anonymized.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, in Appendix A, the dataset description.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Yes, tables reporting statistics can be found in Section 3, problem definition, and appendix A, Dataset Description.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Yes, as shown in Section 5 and Appendix C.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Yes, parameters are reported in Table 11 in Appendices, and computational budgets are reported in Appendix C.1.2 and Appendix C.2.1.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, in Appendix C.1.2 and Appendix C.2.1.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Yes, in Section 5 and Appendix C.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Yes, the usage of these packages is well-explained, and packages are cited throughout the Appendix.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Yes, expert annotations are conducted in Section 5.2, Generative Event Conceptualization. Details are introduced in Appendix C.2.2.

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Yes, in Appendix C.2.2.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Yes, in Appendix C.2.2.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Yes, in Appendix C.2.2.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Not applicable.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Not applicable.