Symbolic Knowledge Distillation:
from General Language Models to Commonsense Models

Peter West[†‡*]   **Chandra Bhagavatula**[‡]   **Jack Hessel**[‡]   **Jena D. Hwang**[‡]
Liwei Jiang[†‡]   **Ronan Le Bras**[‡]   **Ximing Lu**[†‡]   **Sean Welleck**[†‡]   **Yejin Choi**[†‡*]

[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence

Abstract

The common practice for training commonsense models has gone *from–human–to–corpus–to–machine*: humans author commonsense knowledge graphs in order to train commonsense models. In this work, we investigate an alternative, *from–machine–to–corpus–to–machine*: general language models author these commonsense knowledge graphs to train commonsense models.

Our study leads to a new framework, **Symbolic Knowledge Distillation**. As with prior art in Knowledge Distillation (Hinton et al., 2015), our approach uses larger models to teach smaller models. A key difference is that we distill knowledge symbolically–as text–in addition to the neural model. We also distill only one aspect–the commonsense of a general language model teacher, allowing the student to be a different type, a commonsense model. Altogether, we show that careful prompt engineering and a separately trained critic model allow us to selectively distill high-quality causal commonsense from GPT-3, a general language model.

Empirical results demonstrate that, for the first time, a *human-authored* commonsense knowledge graph is surpassed by our *automatically distilled* variant in all three criteria: quantity, quality, and diversity. In addition, it results in a neural commonsense model that surpasses the teacher model's commonsense capabilities despite its 100x smaller size. We apply this to the ATOMIC resource, and share our new symbolic knowledge graph and commonsense models[1].

1 Introduction

Prior works have suggested that pre-trained neural language models have limited understanding of commonsense knowledge (Merrill et al., 2021;
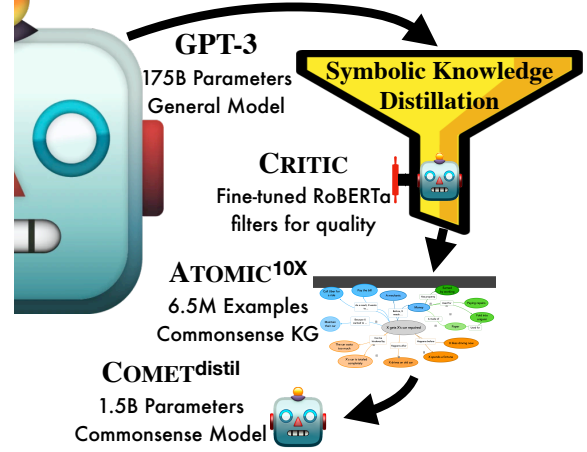
---

[1]http://www.symbolicKD.com/

Figure 1: **Symbolic knowledge distillation** extracts the commonsense from the large, general language model GPT-3, into 2 forms: a large commonsense knowledge graph ATOMIC[10x], and a compact commonsense model COMET$_{\text{TIL}}^{\text{DIS}}$. The quality of this knowledge can be controlled and improved by adding a **critic** model, making GPT-3 a stronger teacher.

Talmor et al., 2021; Davis and Marcus, 2017) despite their otherwise stellar performances on leaderboards. As a result, symbolic commonsense knowledge graphs (Speer et al., 2017; Sap et al., 2019; Hwang et al., 2020) and their corresponding neural representations (Bosselut et al., 2019; Hwang et al., 2020; Zhang et al., 2020b) have been useful for supplementing past models with commonsense capabilities for diverse downstream applications, including interactive learning through a conversational interface (Arabshahi et al., 2021), persona- and affect-aware conversation models (Kearns et al., 2020), figurative language understanding (Chakrabarty et al., 2020, 2021), story telling (Ammanabrolu et al., 2020) and fantasy games (Ammanabrolu et al., 2021).

The common practice for constructing commonsense knowledge graphs has been asking humans to spell out as many pieces of commonsense knowledge as possible. This corresponds to a pipeline

from–human–to–corpus–to–machine, where the commonsense models are trained from human-authored knowledge graphs. Yet, high-quality, human-authored knowledge graphs are too expensive to scale, limiting coverage, which motivates an alternative: *from–machine–to–corpus–to–machine*. Prior efforts toward automatically constructed commonsense knowledge graphs have shown considerably lower quality than humans (Hwang et al., 2020; Zhang et al., 2020b), which in turn leads to neural models that perform less reliably (Hwang et al., 2020). In fact, the broader literature on automatic knowledge graph construction has shown, time and again, that machine-authored knowledge graphs do not match the quality of human-authored knowledge graphs (Etzioni et al., 2011; Mitchell et al., 2015; Bollacker et al., 2008).

In this work, we propose **Symbolic knowledge distillation** as a new conceptual framework toward high-quality automatic knowledge graphs for commonsense, leveraging state-of-the-art models and novel methodology. Most prior art for automatic knowledge graph construction has extracted knowledge from raw text (Bhakthavatsalam et al., 2020; Zhang et al., 2020a; Zhou et al., 2020; Zhang et al., 2020b; Li et al., 2020). In contrast, our approach is motivated by earlier work in Knowledge Distillation (Hinton et al., 2015) wherein a larger teacher model transfers knowledge to a more compact student model (§2.1). Key differences compared to prior knowledge distillation are that we distill a symbolic knowledge graph (i.e., generated text) in addition to a neural model, and that we distill only a selective aspect of the teacher model. This selective distillation allows for the student model to be of a different type (a commonsense model), compared to the teacher (a general language model), thus the possible scope of distillation is enriched. The added benefit of distilling knowledge in a symbolic form is that it is human readable: it can be understood and evaluated.

On the other hand, going from a general language model to specialized commonsense model poses a unique challenge, as the teacher model's knowledge, off-the-self, is inevitably noisy. This is in line with prior work, showing mixed success with generated knowledge. A general language model–GPT-3 in our case–is an imperfect commonsense teacher on its own. Yet we empirically demonstrate that, with a separately-trained critic model judging quality, a more precise and critical

teacher can be defined. Knowledge from the critical teacher can be higher quality–even exceeding that of human-authored knowledge. Furthermore, our study leads to the unexpected finding that even without a critical teacher, the student model is able to surpass the commonsense of GPT-3, our knowledge source.

To test symbolic knowledge distillation against the *human-to-corpus-to-model* paradigm, we compare to ATOMIC_{20}^{20} (Hwang et al., 2020), which is a human-authored commonsense knowledge graph. We find that **ATOMIC^{10x}**, our machine-generated corpus, exceeds the human generated corpus in *scale, accuracy,* and *diversity* with respect to 7 commonsense inference types that we focus on in this study. The resulting commonsense model, **COMET$_{\text{TIL}}^{\text{DIS}}$**, not only surpasses the human-trained equivalent COMET_{20}^{20}, but is also smaller, more efficient, and produces commonsense at a higher accuracy than its own teacher–GPT-3.

Symbolic knowledge distillation offers a promising new role for general language models, as commonsense knowledge sources. It also offers a new role for humans, as small-scale evaluators to train critic models, rather than knowledge authors. Our work demonstrates that humans and LMs can be effective collaborators for curating commonsense knowledge graphs and training efficient and performant commonsense models.

2 Overview and Key Findings

Throughout our work, we describe the *model–to–corpus–to–model* methodology of symbolic knowledge distillation, starting with a teacher model and ending with a commonsense model. We start by describing *model–to–corpus*, from a *loose teacher*, GPT-3 only (§3). Under a fixed budget, this produces large-scale yet noisy knowledge. Next, we improve this by introducing a critic model, resulting in a *critical teacher*. With the same budget, this produces smaller scale but higher quality knowledge (§4). Finally, we go from *corpus–to–model*, using knowledge generated by different teachers to train specialized commonsense models. Throughout, we contrast past work using humans as a knowledge source, comparing the human-authored commonsense knowledge graph ATOMIC_{20}^{20} and resulting commonsense model COMET_{20}^{20} (Hwang et al., 2020) to our own **ATOMIC^{10x}** and **COMET$_{\text{TIL}}^{\text{DIS}}$**.

Symbolic knowledge distillation naturally fol-

X starts running	**xEffect** *so, X*	gets in shape	X sings a song	**HinderedBy** *but not if*	X can't remember the lyrics
X and Y engage in an argument	**xWant** *so, X wants*	to avoid Y	X is not well liked	**xReact** *so, X feels*	lonely
X learns to type fast	**xNeed** *X needed*	to have taken typing lessons	X takes care of a monkey	**xAttr** *X is seen as*	kind
X steals his grandfather's sword	**xEffect** *so, X*	is punished by his grandfather	X butts in	**HinderedBy** *but not if*	X is too shy to speak up
X takes up new employment	**xIntent** *because X wants*	to be self sufficient	X waits for the storm to break	**xEffect** *so, X*	is safe from the storm

Figure 2: Examples of **automatically generated** ATOMIC triples from our ATOMIC$^{\textbf{10x}}$ commonsense knowledge graph. For each example we include a generated event, corresponding relation (and natural language interpretation), and the resulting generated inference.

lows the large–to–small model compression of knowledge distillation (Hinton et al., 2015), yet our use of generated knowledge results in a fundamentally different process.

2.1 Symbolic Knowledge Distillation

Our proposed methodology parallels knowledge distillation (Hinton et al., 2015), a method for compressing a large or complicated teacher distribution P_t into a smaller/simpler student distribution P_s. Key to knowledge distillation[2] is the notion of minimizing the cross-entropy between P_t and P_s:

$$H(P_t, P_s) = - \sum_{y \in Y} P_t(y) \log P_s(y) \qquad (1)$$

In effect, knowledge is transferred to the student by encouraging it to match predictions of the teacher. Hinton et al. (2015) apply this to conditional classification: for each input example in a training set, P_t and P_s are model predictions over label set Y. Typically Y will be a finite set of tractable size, over which the sum in Eq 1 can be reasonably calculated.

For distilling the knowledge of generative models, we can think of an unconditional language model (LM) (e.g. GPT-3) as P_t. This makes Y the set of all strings, over which LMs define probability. Unfortunately Y is an exponential set, intractable to sum over in Eq 1. Kim and Rush (2016) address this problem by simply taking the mode of P_t over Y, truncating most of the teacher distribution and losing information.

Instead, we consider a sampling-based interpretation of the same objective:

$$H(P_t, P_s) = \mathop{\mathbb{E}}_{y \sim P_t(y)} \left[- \log P_s(y) \right] \qquad (2)$$

which exactly equals the cross-entropy of Eq 1, at the limit under pure sampling from P_t.

Yet our work is focused on commonsense–we do not want *all knowledge* from GPT-3, only this domain. In fact, the true teacher P_t that we would like is a commonsense expert, which GPT-3 is not, off-the-shelf, but we approximate such a teacher via careful prompting. In fact, this is an explicit benefit of the sampling-based interpretation of Eq 2: while Eq 1 used continuous logits, we are working with discrete strings, which allows us to make discrete decisions about which knowledge is transferred.

While GPT-3 is not specifically trained to output commonsense, one simple way to encourage quality in the transferred knowledge is using truncated sampling, such as top-p (Holtzman et al., 2019). This is on top of prompting to encourage the correct domain of knowledge. In this case P_t is the truncated distribution of GPT-3, prompted to return commonsense. We call this the *loose teacher* P_t^L– knowledge is being generated and transferred from GPT-3, but without critical assessment of what is correct. We experiment with this in §3.

In fact, sampling knowledge in Eq 2 offers even more opportunity for more control, as generations can be individually interpreted and judged. If we have an indicator function $A(x)$ to tell us which knowledge x is *high quality*, then we can define a better teacher. Using a Product of Experts (Hinton, 2002) between the loose teacher P_t^L and and the

---

[2]In its simplest case, with temperature set to 1.0

critic $A(x)$, we can define a *critical teacher*:

$$P_t(x) \propto P_t^L(x|p) \cdot A(x) \qquad (3)$$

In practice, $A(x)$ is a textual classifier learned on human judgements, set as 1 for knowledge predicted to be high quality and 0 otherwise. Thus, the critical teacher has an explicit control on the correctness of knowledge it transfers. We explore this option in §4. While the critical teacher results in higher quality knowledge, per-example computational cost is higher as each one is sampled from P_t, then accepted or rejected by A.

2.2 Key Findings

Applying Symbolic knowledge distillation in practice results in a number of promising and surprising findings:

1. Learning symbolic knowledge from language models can be framed as a symbolic extension to knowledge distillation. In §2.1, we describe learning commonsense as a symbolic extension to knowledge distillation with GPT-3 as the teacher to smaller models. We elaborate on this process and positive results in §3, 4, and 5.

2. Symbolic knowledge distillation constructs a high quality knowledge graph at scale. A natural byproduct of our method is a machine-generated commonsense knowledge graph. In §4, we show how this graph can achieve impressive quality. The key to success is an effective critic model to filter out potentially incorrect generated knowledge.

3. A critical teacher results in a higher quality student. In §4, we show that making the teacher more critical results in higher quality knowledge, even as it reduces the scale of knowledge transferred. This demonstrates that *quality* matters, not just *quantity*, as higher quality knowledge results in a higher quality commonsense model in §5.

4. Critical teachers or not, a student can outperform the knowledge source. In §5, we show the unexpected result, that all student models exceed the quality of GPT-3, the basis of our teacher distributions.

5. Machines can now win over humans for automatic knowledge graph construction. In §4 and §5, we show that machine generated knowledge and the resulting commonsense model do better than using a human knowledge source. Our symbolic knowledge exceeds humans at scale, quality,

and diversity. Our resulting commonsense model achieves the highest accuracy commonsense KG completions.

3 Model-to-Corpus Verbalization

Symbolic knowledge distillation begins by going *model-to-corpus*, i.e., verbalizing a large number of commonsense facts in a natural language format. This results in a commonsense knowledge graph. From §2.1, we can see this as taking samples to estimate the knowledge distillation objective, encouraging a student model to match a teacher distribution. By generating, GPT-3 teaches a smaller commonsense model trained on those generations.

We start with a *loose teacher*, which simply transfers knowledge by prompted generation with truncated sampling–this is in contrast to the *critical teacher* of §4. The loose teacher follows a few-shot prompting paradigm from Brown et al. (2020). We condition using this template:

```
<TASK-PROMPT>
<EX₁-INP><EX₁-OUT>
<EX₂-INP><EX₂-OUT>
...
<EX_{N-1}-INP><EX_{N-1}-OUT>
<EX_N-INP>
```

where $<\text{EX}_i\text{-INP}>/<\text{EX}_i\text{-OUT}>$ are a small number of human-authored, natural language linearizations of ATOMIC entires, and `<TASK-PROMPT>` is an optional natural language description of the completion problem. Given a prompt of this format, GPT-3 generates the *missing piece*, i.e., the output $<\text{EX}_N\text{-OUT}>$ for input $<\text{EX}_N\text{-INP}>$, following the pattern of earlier examples (1 to N-1). While verbalization largely follows prior work, we find a number of important aspects for producing high-quality commonsense knowledge:

- Examples should be numbered. For instance, $<\text{EX}_5\text{-INP}>$ might begin with "5)" to indicate it is the 5th example. This increases the degree to which GPT-3 follows previous examples.
- The format of $<\text{EX}_i\text{-INP}>$ and $<\text{EX}_i\text{-OUT}>$ should linguistically imply the relationship between them. See below for examples.
- `<TASK-PROMPT>` can be used to give extra specification to complicated problems.

3.1 Data: ATOMIC

We demonstrate symbolic knowledge distillation on the ATOMIC *if-then* resource (Sap et al., 2019).

ATOMIC follows an event-relation-inference (triple) format. A large set of *events* forms the foundation of the corpus, e.g. *X attacks Y*. An ATOMIC triple links a relation to an event, like the **HinderedBy** relation which describes something that might hinder this event. For a given relation and event, the goal is to generate a resulting inference, e.g. *X attacks Y* **HinderedBy** *X is restrained*.

Of the 23 relations from the most recent commonsense knowledge graph–ATOMIC$_{20}^{20}$–we limit our investigation to 7 relations that correspond to *causal* commonsense knowledge[3]: **xAttr** (how X is perceived after *event*), **xReact** (how X reacts in response to *event*), **xEffect** (what X does after *event*), **xIntent** (X's intent in *event*), **xWant** (what X wants after *event*), **xNeed** (what X needed for *event* to happen) and **HinderedBy**.

We describe how **verbalization** is applied to ATOMIC data in 2 steps: generating the underlying events (heads), then generating full examples (inference given event).

3.2 Event Generation

Events are context-free premises in ATOMIC involving `PersonX` (and sometimes a second `PersonY`) in various scenarios; these events form heads in the knowledge graph triples. We generate events by filling in the elements of our template:

```
1.  Event:  X overcomes evil with good
2.  Event:  X does not learn from Y
…
10.  Event:  X looks at flowers
11.
```

following a fairly simple format, as events are generated *unconditionally*.

We collect a set of 100 high-quality events from ATOMIC$_{20}^{20}$ to use in our prompt, randomly sampling 10 each time we generate. We use nucleus sampling with $p = 0.9$ (Holtzman et al., 2019), and presence/frequency penalties of 0.5 from the GPT-3 interface. We generate 165K unique events using the 175B-parameter Davinci model[4] from Brown et al. (2020) (human-authored ATOMIC$_{20}^{20}$ contains 6.2K events–generation greatly increases scale).

3.3 Inference Generation

Generating ATOMIC inferences requires conditioning on events and relations, and requires reasoning about logical relationships between premise events and conclusions. We found that both describing the task and linearizing input/output pairs with natural language connections greatly increases quality. For each relation, we design a custom verbalization format, using iterative design and small-scale verification by the authors.[5] For example, consider the **xNeed** relation, prompted using our template as follows:

```
What needs to be true for this
event to take place?

...

Event <i>:  X goes jogging

Prerequisites:  For this to
happen, X needed to wear running
shoes

...

Event <i>:  X looks at flowers

Prerequisites:  For this to
happen,
```

We use natural language that implies the generation should be a requirement for the input event by prefacing the output with "Prerequisites:" and beginning the sentence with "for this to happen". We also include an xNeed-specific `<TASK-PROMPT>`. As in §3.2, we sample 10 few-shot examples for each prompt from a set of 100 human-authored cases.[6]

For each pair of event (165K) and relation (7) we generate 10 inferences with the second largest form of GPT-3, Curie[7], following the same hyperparameters as event generation. After removing duplicate and degenerate (e.g. fewer than 3 characters) generations, this results in 6.46M ATOMIC-style data triples. We call this ATOMIC**10x**, as it contains an order of magnitude more triples than ATOMIC$_{20}^{20}$ with respect to the 7 relations we investigate. Table 1 compares the magnitude and cost-of-creation for both corpora.

3.4 Evaluating a Generated Commonsense Knowledge Graph

While machine generation enables a much larger scale of unique generations, what kind of knowl-

---

edge is produced by GPT-3, and how does it differ from knowledge produced by humans? We can start to understand the attributes by looking at the generated triples in Fig 2, but we continue with more in-depth analysis.

Lexical Differences: Diversity and Uniqueness Recent work has found that machine generation can be highly repetitive or lack diversity (Welleck et al., 2020; Holtzman et al., 2019); one way we might expect generated knowledge to differ from human authored is in less creative word choice, less diversity, or more repetition. Is this actually the case?

We begin simply with lexical diversity, i.e., the number of unique words used (Table 2). While diversity differs by relation, the machine generated ATOMIC[10x] demonstrates high lexical diversity across both events and inferences, using 881K unique words vs. 109K in human-authored ATOMIC$_{20}^{20}$.

In addition to containing many strictly unique generated inferences (Figure 3), we also sought to quantify more subtle notions of diversity and novelty.

BLEU Soft Uniqueness. Following a similar intuition to self-BLEU (Zhu et al., 2018), we define a notion of *soft uniqueness* to describe how close or diverse different generations are. We say an inference x is softly-unique if:

$$BLEU_2(C, x) < 0.5$$

where C is the set of inferences generated for the same event/relation input, and 0.5 is an empirically chosen threshold. To summarize the soft-uniqueness of the whole corpus, we iteratively remove examples until the entire corpus is softly unique. This process results in a set of inferences with low mutual lexical overlap; the size of the resulting corpus provides perspective on the diversity of the original corpus (bigger=better). We include the softly-unique sizes of each corpus in Table 4 ("Size (div)"). Although ATOMIC[10x] contains a slightly lower *percentage* of softly-unique points than ATOMIC$_{20}^{20}$, it has a much larger number of these diverse examples, e.g., ATOMIC[10x]'s soft de-duplication results in $4.38M$ entries (starting from $6.5M$) in comparison to ATOMIC$_{20}^{20}$, which has $560K$ soft de-duplicated entries (starting from $600K$).

Relation	ATOMIC$_{20}^{20}$	ATOMIC[10x]
HinderedBy	77,616	**1,028,092**
xNeed	100,995	**760,232**
xWant	109,098	**730,223**
xIntent	54,839	**965,921**
xReact	62,424	**1,033,123**
xAttr	113,096	**884,318**
xEffect	90,868	**1,054,391**
Total Count	608,936	**6,456,300**
Est Total Cost	~$40,000	**~$6,000**
Est Cost Per Triple	~$0.06	**~$0.001**

Table 1: Number of unique triples with the given relation, $|(\cdot, \texttt{relation}, \cdot)|$. The estimated cost for ATOMIC[10x] comes at a fraction of a conservative estimation for ATOMIC$_{20}^{20}$ crowdsourcing costs.

	Length		Unique Tokens	
	AT$_{20}^{20}$	AT[10x]	AT$_{20}^{20}$	AT[10x]
xWant	4.69	5.16	322K	784K
xAttr	1.42	2.73	15K	21K
xEffect	3.92	4.66	216K	864K
xIntent	4.59	5.92	136K	800K
xNeed	4.51	5.97	289K	1,378K
xReact	4.03	1.77	48K	5K
HinderedBy	7.93	7.49	522K	1,775K
Events	5.20	**5.32**	109K	**881K**

Table 2: Average length and total unique tokens in inferences by relation type, and in events (bottom row) from ATOMIC$_{20}^{20}$ (AT$_{20}^{20}$) and ATOMIC[10x] (AT[10x]). EVent statistics are reported once for all relations since they remain essentially constant across all relations.

Model-based Diversity Measurement. As another measurement of information diversity, we use information-theoretic measures. Intuitively, diverse information should be less predictable, thus have higher entropy. Taking GPT-2 XL models finetuned on ATOMIC$_{20}^{20}$ and ATOMIC[10x] (see §5) we can estimate **entropy**–roughly, how difficult it is for a model to capture all information in a corpus, a notion of content. Also, **cross-entropy**–how well a model trained on one corpus describes the other. We include these estimates in Table 3. Entropy is roughly 4 times higher for ATOMIC[10x], suggesting more content from a modelling perspective. Cross entropy from ATOMIC[10x] to ATOMIC$_{20}^{20}$ is 9.31, only 2 points higher than its entropy suggesting ATOMIC$_{20}^{20}$ is easily describable with information from ATOMIC[10x]. In the other direction, we see a cross entropy of 41.48 suggesting much of ATOMIC[10x] is not captured by ATOMIC$_{20}^{20}$–many examples in ATOMIC[10x] are surprising given only information from ATOMIC$_{20}^{20}$.
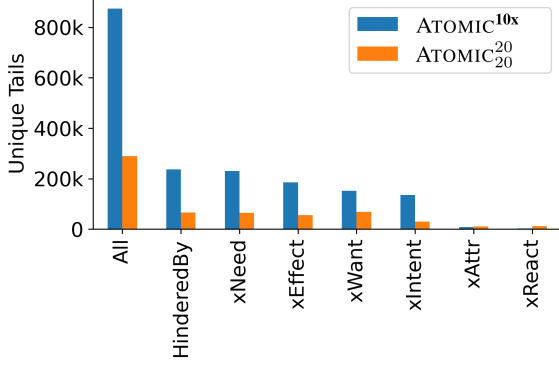
Figure 3: Unique tails (inferences) by relation.

Entropy	Cross Entropy	KL Divergence
$H(D_1) = 1.27$	$H(D_1, D_2) = 9.31$	$D_{KL}(D1||D2) = 8.04$
$H(D_2) = 7.80$	$H(D_2, D_1) = 41.48$	$D_{KL}(D2||D1) = 33.68$

Table 3: Entropy of ATOMIC_{20}^{20} (D_1) and $\text{ATOMIC}^{\mathbf{10x}}$ (D_2), and the cross-entropy and KL divergence between each.

Human Evaluation of Quality. Perhaps most importantly, we study the *quality* of knowledge in each corpus. We conduct a human evaluation using Amazon Mechanical Turk. Workers are presented with ATOMIC-style triples, replacing relations with natural language templates (e.g. *HinderedBy* becomes "can be hindered by"). 3 annotators rate each triple, with options for acceptability: "always/often", "sometimes/likely", "farfetched/never", "invalid", or "too unfamiliar to judge". The first two are considered "accepted", the second two "rejected" and the final is "no judgement". We evaluate 3000 examples from $\text{ATOMIC}^{\mathbf{10x}}$–this ensures at least 1000 evaluated examples after filtering by the critic (§4), which ensures statistically meaningful results in those cases. Results are included in Table 4: For binary acceptability, we find Fleiss' kappa (Fleiss, 1971) of 40.8 indicating moderate agreement (Landis and Koch, 1977). We also fine accuracy agreement of 90.5% over possible annotator comparisons.

For the *loose teacher*, we should only consider the top row of $\text{ATOMIC}^{\mathbf{10x}}$ in Table 4, all other rows add the critic model (§4). $\text{ATOMIC}^{\mathbf{10x}}$ exceeds ATOMIC_{20}^{20} in scale, yet it is somewhat less acceptable by human raters–by roughly 8 percentage points. Yet, while accuracy of $\text{ATOMIC}^{\mathbf{10x}}$ is lower, its much larger scale implies a significantly higher *number* of accurate examples. Increasing the ratio of these high-quality examples within the

Corpus	Accept	Reject	N/A	Size	Size (div)
ATOMIC_{20}^{20}	86.8	11.3	1.9	0.6M	0.56
$\text{ATOMIC}^{\mathbf{10x}}$	78.5	18.7	2.8	**6.5M**	**4.38**
	88.4	9.5	2.1	5.1M	3.68
	91.5	6.8	1.7	4.4M	3.25
	94.3	4.6	1.1	3.6M	2.74
	95.3	3.8	1.0	3.0M	2.33
	96.4	2.7	0.8	2.5M	2.00

Table 4: Attributes of $\text{ATOMIC}^{\mathbf{10x}}$ and $\text{ATOMIC}^{\mathbf{10x}}$ (row 2) including the critic model (§4, rows 3 - 7) with various filtering cutoffs. For human judgements, we include Accept and Reject by majority vote unless at least 1 reviewer marked "too unfamiliar to judge" (N/A). Size is given in millions of unique examples[8]. The highest precision corpus is $\text{AUTOTOMIC}_{0.5}$ but precision of ATOMIC_{20}^{20} still can be exceeded with the much larger $\text{AUTOTOMIC}_{0.9}$. We also include number of diverse examples "Size (div)" described in §3.4.

corpus is the main objective of a critic model (§4).

4 Making the Teacher More Critical

While the *loose teacher*, GPT-3, alone results in a viable commonsense knowledge graph, our human evaluations demonstrate that this isn't a perfect commonsense teacher. We consider multiplying in a *critic model*, which aims to filter out lower-quality knowledge, thereby *correcting the teacher* (§2.1). With a modest amount of supervision in the form of a small-scale human evaluation, we can train a classifier to predict and discriminate unacceptable examples. We can imagine multiplying this with the *loose teacher* §3, create a *critical teacher* product of experts. The critic removes unacceptable knowledge; in practice this means filtering the generations in $\text{ATOMIC}^{\mathbf{10x}}$ and creating a range of new corpora that are higher quality, yet still larger scale than ATOMIC_{20}^{20}.

We first gather a training set of *correct vs. incorrect* human judgments on a randomly-sampled set of 10K entries of $\text{ATOMIC}^{\mathbf{10x}}$. We follow the evaluation setup from § 3.4, but gather only one annotation for each example. We take a (random) train/dev/test split of 8k/1k/1k. While this aspect of symbolic knowledge distillation is not entirely automatic, evaluating 10k examples is both simpler and vastly cheaper than human construction of a corpus of equivalent scale. Instead, we argue that a more useful and efficient role for humans in

---

[8]Size of ATOMIC_{20}^{20} is given as the number of comparable datapoints, i.e. those with the same relations as $\text{ATOMIC}^{\mathbf{10x}}$.

knowledge graph construction is to correct the mistakes of the teacher by evaluating a small number of examples.

We train binary classifiers (critics) for human acceptability using RoBERTa-Large (Liu et al., 2019), fine-tuning all parameters, along with a 2-layer MLP on the `[CLF]` representation. We conduct a small grid search on the validation set finding batch size 128, dropout .1, and Adam (Kingma and Ba, 2014) learning rate 5e-6 to be effective. We use early stopping and decay learning rate on validation performance plateauing, to maximize $R@80\%$ on the validation set. We find RoBERTa pretrained on MNLI (Williams et al., 2018) effective, outperforming other options. As well, we substitute randomly-sampled names in for person designations "X"/"Y". We include as a baseline an unsupervised filtration metric inspired by (Davison et al., 2019): they propose a model estimate of PMI to score mined commonsense triples. In our case, we use Negative Log-Likelihood (NLL) and token-mean-NLL from GPT-3 itself.

The validation precision/recall of our best performing model, the baselines, and the in-optimal hyperparameter configurations are given in Figure 4. Once fixing our model, we applied it to the test set (also in Fig 4), verifying that it generalizes to ATOMIC$^{\mathbf{10x}}$ entries. Overall, our trained critic model is more effective than the baselines in identifying high and low quality teacher generations at all levels of precision and recall. This result demonstrates that a small amount of human supervision can consistently help to correct GPT-3's mistakes.

Size-accuracy Trade-off As with any filtration process, our critic faces a trade-off between size and accuracy based on what cutoff is used within the classifier to make a final filtration decision. We present several versions of ATOMIC$^{\mathbf{10x}}$ representing different cutoffs, i.e. the confidence at which the critic rejects a generated example. To verify that our critic is working as expected, we perform an additional human validation for each cutoff by annotating an additional, results are in the Accept/Reject column of Table 4.

By discarding the 20% of the instances the classifier judges as least acceptable (reducing knowledge graph size from 6.5M to 5.1M) ATOMIC$^{\mathbf{10x}}$'s accuracy rises from $78.5 \rightarrow 88.4$; human-authored ATOMIC$^{20}_{20}$ contains only 600K entries at 86.8% accuracy. Reducing to 2.5M examples (38% of the original size), we can attain an accuracy of 96.4%,
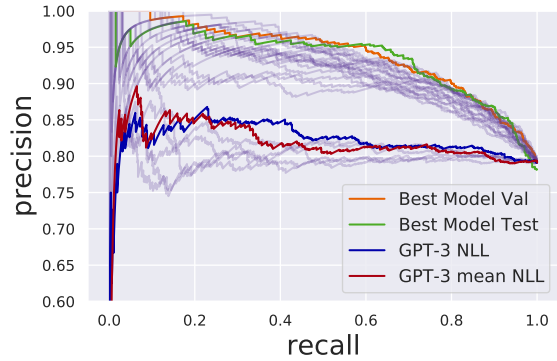

Figure 4: Precision vs. recall of our critic model on the human labelled validation set. The best trained models are labelled, and other hyper-parameter settings are shown as faded lines. We also include generation negative log-likelihood (nll) and token-wise mean nll as cutoff measures–these perform much worse than the supervised model.

nearly 10 points above the human corpus. To put this in perspective, at this cutoff, ATOMIC$^{\mathbf{10x}}$ is still 4X larger than ATOMIC$^{20}_{20}$.

What gets filtered out? We qualitatively identify two types of filtered triples: 1) *logical misalignments*, which consist of events/inferences joined in a logically inconsistent manner. Identification of these triples requires modeling logical interactions between events and inferences, e.g., *X cannot find his shirt **as a result X is wearing a shirt**, X was kidnapped **as a result X feels** safe.*; and 2) *awkward phrasings*, which consist of events/inferences that, in isolation, are incoherent, ambiguous, or awkwardly phrased. Examples of events with awkward phrasings that were include: *PersonX has a fire in the bath* and *PersonX appears to be three-dimensional*; any inference that completes these events may be unacceptable because the event itself is broadly implausible.

To understand what types of triples are filtered (and in what frequency), we ablate our critic model: the validation precision-recall results, summarized by average precision (AP), are given in Table 5: our full model is compared to a random predictor, a event-only model, and a inference-only model, all based on RoBERTa. We also compare to an additive EMAP (Hessel and Lee, 2020) version of our model that computes an optimal additive ensemble of event-only and inference-only models, but doesn't allow for logical interactions between the event/inference to be incorporated (as would be required to identify *logical misalignments*).

	AP
Random	79.3
Inference-only	81.9
Event-only	86.2
EMAP(Full model)	87.1
Full model	**94.0**

Table 5: Average Precision for ablations of the critic model (higher=better). The critic model not only filters out *awkward phrasings* (which can be identified looking at either the event or inference in isolation: EMAP can only identify these), but also filters out *logical misalignments*, which require modeling interactions between the event/inference.

We find that GPT-3 produces both independently awkwardly-phrased events/inferences and also logical misalignments. But, our classifier, trained on validated knowledge triples, helps in both cases. The EMAP of our full model (which can identify only awkward phrasings) achieves 87% AP, and our full model (which can additionally identify logical misalignments) improves performance to 94% AP.

Does Filtering Hurt Diversity? One could imagine the critic reducing ATOMIC$^{\textbf{10x}}$ to only very similar "safe" examples that might not be as useful/interesting in practice. We repeated our diversity analysis from §3.4 for the various cutoffs; results are in Table 4 ("Size (div)", higher=better). As we filter more stringently, we surprisingly observe proportionally *more* diverse examples: full size ATOMIC$^{\textbf{10x}}$ has a diverse subset which is 68% of its original size, whereas its most filtered version goes up to 80%. One possible reason is if GPT-3 gravitates towards some common sentence structures that represent inconsistent knowledge. These would be highly recognizable across event/relation pairs, and removing them would increase both quality and diversity. This surprising result warrants further study.

5   Corpus-to-Model: Distillation

The final step of symbolic knowledge distillation is to train a compact model on the extracted natural language knowledge. Our base model is GPT2-XL trained on all of ATOMIC$^{\textbf{10x}}$: we denote this model by COMET$_{\text{TIL}}^{\text{DIS}}$. We additionally consider training the same model on the critical versions of ATOMIC$^{\textbf{10x}}$–crit$_{\text{low}}$ denotes training on the 91.5%

CKG Completion Model	Train Corpus Acc	Accept	Reject	N/A
GPT2-XL zero-shot	-	45.1	50.3	4.6
GPT-3	-	73.3	24.1	2.6
COMET$_{20}^{20}$	86.8	81.5	16.3	2.2
COMET$_{\text{TIL}}^{\text{DIS}}$	78.5	78.4	19.2	2.4
+critic$_{\text{low}}$	91.5	82.9	14.9	2.2
+critic$_{\text{high}}$	96.4	**87.5**	10.2	2.3

Table 6: Human judgements for knowledge-base completion on held out events from the ATOMIC$_{20}^{20}$ commonsense knowledge graph (CKG). We test GPT-2 XL models–both a zero-shot baseline and COMET models trained on ATOMIC$_{20}^{20}$ (COMET$_{20}^{20}$) and ATOMIC$^{\textbf{10x}}$ (COMET$_{\text{TIL}}^{\text{DIS}}$). We include 3 settings for COMET$_{\text{TIL}}^{\text{DIS}}$: first trained on unfiltered ATOMIC$^{\textbf{10x}}$, then filtered by the critic to 88.4% accuracy (+critic$_{\text{low}}$) and 96.4% accuracy (+critic$_{\text{high}}$). We find the models trained on ATOMIC$^{\textbf{10x}}$ have the highest acceptance rates, and the model trained on the most filtered version performs best. We also include GPT-3 as a baseline using the same technique as §3, which performs significantly worse than COMET$_{\text{TIL}}^{\text{DIS}}$ trained on this style of data.

accuracy corpus, and crit$_{\text{high}}$ on the 96.4% version. All models are trained for 1 epoch, with default parameters using the Huggingface Transformers library (Wolf et al., 2019).

5.1   Evaluating a Symbolically Distilled Model

Evaluation follows past work (Hwang et al., 2020; Bosselut et al., 2019; Sap et al., 2019) and tests the ability of models to do automatic knowledge base completion. COMET$_{\text{TIL}}^{\text{DIS}}$ and baselines must generate inferences for events that did not appear in its training set. We carry out generation on a subset of the ATOMIC$_{20}^{20}$ test set, and evaluate generations by human evaluation[9] following the evaluation setup from Section 3.4. We carry this out on 1000 inputs (event + relation) with results in Table 6. We compare to the GPT2-XL-based COMET$_{20}^{20}$ model, which is trained on the human-generated ATOMIC$_{20}^{20}$. We compare to GPT-3, using the same generation method as verbalization in §3– in effect, we are comparing the student COMET$_{\text{TIL}}^{\text{DIS}}$ to the *loose teacher* GPT-3. Note that the *critical teacher* is not assured to produce inferences for every event/relation pair: it is possible that the critic will reject nearly all tails for some inputs, making

---

[9]We find Fleiss' kappa (Fleiss, 1971) of 47.1 for acceptance, indicating moderate agreement. (Landis and Koch, 1977). We also find agreement of 88.7% over possible annotator comparisons.

this an infeasible baseline. Finally, we compare to zero-shot GPT-XL (Radford et al., 2019) using this same methodology. Results are presented in Table 6.

How does $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ compare to GPT-3? In knowledge distillation, the student model is generally expected to lose some performance (Hinton et al., 2015; Kim and Rush, 2016) compared to its teacher. Thus, we can first compare the base teacher–GPT-3–to the simplest version of $\text{COMET}_{\text{TIL}}^{\text{DIS}}$, without the critic; this is the top-row $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ of Table 6. Surprisingly, it surpasses GPT-3, the model that generated its training data, using the same verbalization technique[10]. We posit that the superior performance of $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ may have to do with mistakes of GPT-3 being filtered out in the verbalization and training process of tuning GPT-2, and possibly the focus of $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ on one commonsense domain while GPT-3 must be more general. However, we leave further study of this effect for future work.

How does $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ compare to human knowledge? While $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ *without the critic* is slightly outperformed by COMET_{20}^{20} in terms of accuracy, this reverses when the critic is introduced. For both cutoffs tested, $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ surpasses COMET_{20}^{20}, with more filtering resulting in a wider gap.

Usefulness of $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ Beyond exceeding commonsense models of the same scale, $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ has concrete utility. For on-demand commonsense inference: given an event and a relation, suppose a high quality inference is needed. In this case, $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ is actually the **best available model**: our $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ trained on the most filtered corpus surpasses COMET_{20}^{20} by 5 points and GPT-3 by over 10. While GPT-3 with the critic can result in a higher acceptability *corpus*, computational overhead aside, the teacher model does not work as reliably for on-demand generation of single examples. For automatic completion of new event/relation inputs, $\text{COMET}_{\text{TIL}}^{\text{DIS}}$ trained on the most critical corpus is the most reliable option.

---

[10]The slight difference in acceptability for GPT-3 from Table 4 is likely due to variance in raters between different rounds of evaluation, as well as a different distribution of events–Table 4 uses generated events while Table 6 uses held out events from ATOMIC_{20}^{20}.

6 Related Work

Commonsense Knowledge Graphs (CKG) Large-scale CKGs provide everyday knowledge for commonsense reasoning. Some are manually constructed, such as ATOMIC (Sap et al., 2019) (880K) whose *if-then* format we follow here. ATOMIC_{20}^{20} (Hwang et al., 2020) extends ATOMIC_{20}^{20} to more examples and relations. ConceptNet (Speer et al., 2017) focuses on taxonomic knowledge and physical commonsense, much of which is authored by humans or compiled from such sources.

Among automatically constructed CKGs are TransOMCS (Zhang et al., 2020a) and CausalBank (Li et al., 2020). TransOMCS contains 18.48M tuples automatically extracted from syntactic parses of various data sources (e.g. Wikipedia) while CausalBank contains 314M cause-effect pairs extracted via pattern-matching from Common Crawl Corpus (Buck et al., 2014). Unlike these works, we *generate* commonsense from a language model.

Extracting Knowledge from LMs Some past work does automatic knowledge base completion with trained models (Bosselut et al., 2019; Hwang et al., 2020; Li et al., 2020), yet these models are trained on *existing* resources of the same format; ATOMIC[10x] is generated from scratch. Other works attempt to mine relational factual and commonsense knowledge directly from off-the-shelf LMs (Petroni et al., 2019; Davison et al., 2019; Xiong et al., 2020), but not resulting in the quality at scale of ATOMIC[10x]. The recent rise of massive pretrained LMs such as GPT-3 (Brown et al., 2020), achieving huge success in downstream NLP tasks and superior in-context few-shot learning ability (Hendrycks et al., 2021) makes these clear sources for commonsense.

Knowledge Distillation Symbolic knowledge distillation parallels knowledge distillation Hinton et al. (2015), which transfers knowledge from a teacher to student distribution. Some work transfers this notion to generation by label smoothing (Sanh et al., 2019), while we distil through generation. Kim and Rush (2016) follow a similar formulation as us, but use the mode of the teacher distribution rather than sampling. One major divergence from similar work is that we distill *specific* information (commonsense) from a general language model into a specialized downstream model.

Data Generation Although manual creation of datasets is expensive and introduces annotation artifacts (Schwartz et al., 2017; Gururangan et al., 2018; Agrawal et al., 2018; Tsuchiya, 2018; Geva et al., 2019; Bras et al., 2020), crowdsourcing remains the most popular method for curating goal-oriented datasets with high coverage and quality.

Past automatic data curation mainly focuses on extractive approaches such as syntactic parsing (Zhang et al., 2020a) and pattern matching (Li et al., 2020) from existing free-form text resources, (e.g. Wikipedia (Lehmann et al., 2015), web crawl (Buck et al., 2014)). Extractive methods are easy to scale but often subject to noise. They are also limited in format–most ATOMIC knowledge will not appear in natural text in an easy-to-mine format. Rather, our work frames GPT-3 as aggregating such textual knowledge into a queryable neural form. With the ever-growing size and capability of pretrained language models (PLMs), some recent works explore automatically dataset synthesis and expansion by finetuning PLMs on the existing labeled data (Anaby-Tavor et al., 2020; Papanikolaou and Pierleoni, 2020; Kumar et al., 2020; Yang et al., 2020). These methods are often limited by the quality of the supervised source data and the capability of relatively small generator PLMs.Schick and Schütze (2021) explores unsupervised generation of datasets for semantically related sentence pairs by giving GPT2-XL instructions. Although the machine generated dataset results in improvements over certain downstream tasks, it is unclear how it compares to human annotation.

7   Conclusions

We introduce symbolic knowledge distillation, a *model-to-corpus-to-model* pipeline for commonsense that does not require human-authored knowledge–instead, using machine generation. We transfer knowledge from a large, general model to a more compact commonsense model, through a commonsense corpus–this results in both a high quality commonsense knowledge graph and model. The resulting symbolic knowledge graph has greater scale and diversity than human authoring, and can exceed its quality as well. Overall, symbolic knowledge distillation offers a high-quality alternative to human-authored knowledge in commonsense research.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.

Prithviraj Ammanabrolu, W. Cheung, William Broniec, and M. Riedl. 2020. Automated storytelling via causal, commonsense plot ordering. *ArXiv*, abs/2009.00829.

Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur D. Szlam, Tim Rocktaschel, and Jason Weston. 2021. How to motivate your dragon: Teaching goal-driven agents to speak and act in fantasy worlds. In *NAACL*.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7383–7390.

Forough Arabshahi, Jennifer Lee, Antoine Bosselut, Yejin Choi, and Tom. Mitchell. 2021. Conversational multi-hop reasoning with neural commonsense knowledge and symbolic logic rules. *ArXiv*, abs/2109.08544.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter E. Clark. 2020. Genericskb: A knowledge base of generic statements. *ArXiv*, abs/2005.00660.

Kurt D. Bollacker, Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD Conference*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, A. Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. In *ICML*.

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.

Christian Buck, Kenneth Heafield, and Bas Van Ooyen. 2014. N-gram counts and language models from the common crawl. In *LREC*, volume 2, page 4. Citeseer.

Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. 2020. R^3: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *ACL*.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. Mermaid: Metaphor generation with symbolism and discriminative decoding.

Ernest Davis and Gary Marcus. 2017. Causal generative models are just a start. *Behavioral and Brain Sciences*, 40.

Joe Davison, Joshua Feldman, and Alexander M Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, et al. 2011. Open information extraction: The second generation. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

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

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *EMNLP*.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Geoffrey E Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*.

William R. Kearns, Neha Kaura, Myra Divina, Cuong Viet Vo, Dong Si, Teresa M. Ward, and Weichao Yuwen. 2020. A wizard-of-oz interface and persona-based methodology for collecting health counseling dialog. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *EMNLP*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, D. Kontokostas, Pablo N. Mendes, Sebastian Hellmann, M. Morsey, Patrick van Kleef, S. Auer, and C. Bizer. 2015. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6:167–195.

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2020. Guided generation of cause and effect. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*.

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. *arXiv preprint arXiv:1907.11692*.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

Tom Michael Mitchell, William W. Cohen, Estevam R. Hruschka, Partha P. Talukdar, Bo Yang, Justin Betteridge, Andrew Carlson, Bhavana Dalvi, Matt Gardner, Bryan Kisiel, Jayant Krishnamurthy, N. Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil Antonios Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard C. Wang, D. Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. 2015. Never-ending learning. *Communications of the ACM*, 61:103 – 115.

Yannis Papanikolaou and A. Pierleoni. 2020. Dare: Data augmented relation extraction with gpt-2. *ArXiv*, abs/2004.13845.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

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.

Masatoshi Tsuchiya. 2018. Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

S. Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and J. Weston. 2020. Neural text generation with unlikelihood training. *ArXiv*, abs/1908.04319.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *International Conference on Learning Representations*.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.

Hongming Zhang, Daniel Khashabi, Y. Song, and D. Roth. 2020a. TransOMCS: From linguistic graphs to commonsense knowledge. In *IJCAI*.

Hongming Zhang, Xin Liu, Haojie Pan, Y. Song, and C. Leung. 2020b. Aser: A large-scale eventuality knowledge graph. *Proceedings of The Web Conference 2020*.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. *ArXiv*, abs/2005.04304.

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

A Human Evaluation Details

We conduct human evaluations on Amazon Mechanical Turk using the template of Figures 5,6. A rating for when the assertion holds true of always/often or sometimes/likely is considered valid. For the purposes of classification in Section 4, these are considered positive and all other ratings are negative. Further details are available in the main text.

B ATOMIC^{10x} Generation Prompts

We include example prompts for all generations we do, from Table 7 to 14

(WARNING: This HIT may contain adult content. Worker discretion is advised.)

Thanks for participating in this HIT!

> If the data is good, it's good. If bad, then bad. Please annotate as you see not worrying about how many of each label you find yourself assigning! If you understand the words but the Phrases or the complete assertation makes poor sense, please mark as INVALID. Thank you!

You will evaluate how often assertions are true. Each assertion is comprised of 3 parts: *Phrase A*, *Relation*, *Phrase B*

Phrase A*, *Phrase B	Short phrases. May describe objects, object properties, events, actions, etc.
Relation	How *A* relates to *B*.

For each assertion, determine how true it is:

always/often	Always or quite often true.
sometimes/likely	Sometimes is true or true for some people. -or- Likely true.
farfetched/never	False or farfetched, at best. -or- Unlikely to be true.
invalid	This assertion makes no sense (i.e., "what does this even mean?!").
too unfamiliar to judge	Cannot make a fair evaluation. Unfamiliar with one or both of the phrase.

If you see "nothing in particular" for *Phrase B*, assess Phrase B in context:

- Sometimes certain actions can simply be responded to by doing nothing!
- Other times, doing nothing in particular is simply a weird or unlikely reaction to something.
- See examples under tricky relations tagged with [nothing in particular example]

Please report any **prejudiced or inappropriate language**:

- Profane or offensive content (NSFW, R-rated material etc)
- Prejudiced assumptions or derogatory language that <u>villainizes</u> people.
  HOWEVER, please note, not all negative content is derogatory especially if Phrase B is intrinsically what Phrase A means. For example:
  criminals *are characterized by* ***committing crime*** is **OK**.
  ↳ This isn't necessarily villianizing people since "criminal" means "a person who has commited a crime".
  homeless *are characterized by* ***being lazy*** is **prejudiced**.
  ↳ There are many reason a person is rendered homeless. This is a gratuitous prejudice about homelessness.
- Material that people may find disturbing, off-putting, or improper

A couple NOTES:

- Please be **forgiving** of *spelling or grammatical errors*
- If the terms are too obscure or you don't know the truth of the fact at the top of your head, it is okay to mark is "too unfamiliar to judge". If you can answer (e.g., based on likelihood), please provide a response.

Figure 5: Page 1 of template used for human evaluation.

1) **PersonX approaches PersonY's aunt**, as a result, PersonX feels, ***awkward***

How often does the assertion hold true?

always/often	sometimes/likely	farfetched/never	invalid	too unfamiliar to judge

○ This fact is true but outdated

☐ I would count this as an inappropriate, prejudiced or offensive material

---

2) **PersonX asked PersonY out on a date**, can be hindered by, ***PersonX is still dating Sarah***

How often does the assertion hold true?

always/often	sometimes/likely	farfetched/never	invalid	too unfamiliar to judge

○ This fact is true but outdated

☐ I would count this as an inappropriate, prejudiced or offensive material

---

3) **PersonX fails to go home**, as a result, PersonX, ***is grounded***

How often does the assertion hold true?

always/often	sometimes/likely	farfetched/never	invalid	too unfamiliar to judge

○ This fact is true but outdated

☐ I would count this as an inappropriate, prejudiced or offensive material

---

4) **PersonX makes her own clothes**, as a result, PersonX feels, ***artistic***

How often does the assertion hold true?

always/often	sometimes/likely	farfetched/never	invalid	too unfamiliar to judge

○ This fact is true but outdated

☐ I would count this as an inappropriate, prejudiced or offensive material

---

5) **PersonX notices PersonY's response**, can be hindered by, ***PersonX is distracted by the music***

How often does the assertion hold true?

always/often	sometimes/likely	farfetched/never	invalid	too unfamiliar to judge

○ This fact is true but outdated

☐ I would count this as an inappropriate, prejudiced or offensive material

(Optional) Please let us know if anything was unclear, if you experienced any issues, or if you have any other fedback for us.

You must ACCEPT the HIT before you can submit the results.

Figure 6: Page 2 of template used for human evaluation.

1. Event: PersonX unwraps PersonY's hands

2. Event: PersonX overcomes evil with good

3. Event: PersonX is fed up with the present situation

4. Event: PersonX breaks PersonX's back

5. Event: PersonX calls no one

6. Event: PersonX never gets angry

7. Event: PersonX does not learn from PersonY

8. Event: PersonX refuses to touch PersonY's hands

9. Event: PersonX looks at flowers

10. Event: PersonX unloads an atomic bomb

11. Event:

Table 7: Prompt for head generation.

Next, how are people seen in each situation? Examples:

Situation 1: Devin bullies Jean.

Devin is seen as dominant.

Situation 2: Jamie moves to another city.

Jamie is seen as adventurous.

Situation 3: Sydney changes Ryan's mind.

Sydney is seen as influential.

Situation 4: Lindsay writes a story.

Lindsay is seen as creative.

Situation 5: Rowan covers Pat's expenses.

Rowan is seen as wealthy.

Situation 6: Lee takes time off.

Lee is seen as carefree.

Situation 7: Riley advises Noel.

Riley is seen as informed.

Situation 8: Adrian bursts into tears.

Adrian is seen as depressed.

Situation 9: Hunter deals with problems.

Hunter is seen as responsible.

Situation 10: Sam follows Charlie.

Sam is seen as suspicious.

Situation 11: Alex makes Chris wait.

Alex is seen as

Table 8: Prompt for generating xAttr.

Next, what do situations make people do? Examples:

Situation 1: Devin gets a divorce.

As a result, Devin dates someone new.

Situation 2: Jamie lifts weights.

As a result, Jamie has sore muscles.

Situation 3: Sydney takes Ryan to a bar.

As a result, Sydney gets drunk.

Situation 4: Lindsay decides to hire a tutor.

As a result, Lindsay gets better grades.

Situation 5: Rowan buys Pat drinks.

As a result, Rowan is thanked by Pat.

Situation 6: Lee hears bad news.

As a result, Lee begins to cry.

Situation 7: Riley buys a chocolate bar.

As a result, Riley gets change.

Situation 8: Adrian does a lot of work.

As a result, Adrian gets mental fatigue.

Situation 9: Hunter attends a concert.

As a result, Hunter hears a new song.

Situation 10: Sam gets the job done.

As a result, Sam gets more responsibilities.

Situation 11: Alex makes Chris wait.

As a result, Alex

Table 9: Prompt for generating xEffect.

For each situation, describe the intent. Examples:

Situation 1: Devin gets the newspaper.

Devin intends to read the newspaper.

Situation 2: Jamie works all night.

Jamie intends to meet a deadline.

Situation 3: Sydney destroys Ryan.

Sydney intends to punish Ryan.

Situation 4: Lindsay clears her mind.

Lindsay intends to be ready for a new task.

Situation 5: Rowan wants to start a business.

Rowan intends to be self sufficient.

Situation 6: Lee ensures Ali's safety.

Lee intends to be helpful.

Situation 7: Riley buys lottery tickets.

Riley intends to become rich.

Situation 8: Alex makes Chris wait.

Alex intends

Table 10: Prompt for generating xIntent.

Next, we will discuss what people need for certain situations. Examples:

1. Before Devin makes many new friends, Devin has to spend time with people.

2. Before Jamie gets a date, Jamie has to ask someone out.

3. Before Sydney changes Ryan's mind, Sydney has to think of an argument.

4. Before Lindsay gets a job offer, Lindsay has to apply.

5. Before Rowan takes a quick nap, Rowan has to lie down.

6. Before Lee tries to kiss Ali, Lee has to approach Ali.

7. Before Riley rides Noel's skateboard, Riley has to borrow it.

8. Before Adrian eats the food, Adrian has to prepare a meal.

9. Before Hunter watches Netflix, Hunter has to turn on the TV.

10. Before Sam has a baby shower, Sam has to invite some friends.

11. Before Alex makes Chris wait, Alex has

Table 11: Prompt for generating xNeed.

Next, how do people feel in each situation? Examples:

Situation 1: Devin lives with Jean's family.

Devin feels loved.

Situation 2: Jamie expects to win.

Jamie feels excited.

Situation 3: Sydney comes home late.

Sydney feels tired.

Situation 4: Lindsay sees dolphins.

Lindsay feels joyful.

Situation 5: Rowan causes Pat anxiety.

Rowan feels guilty.

Situation 6: Lee goes broke.

Lee feels embarrassed.

Situation 7: Riley has a drink.

Riley feels refreshed.

Situation 8: Adrian has a heart condition.

Adrian feels scared about their health.

Situation 9: Hunter shaves Avery's hair.

Hunter feels helpful.

Situation 10: Sam loses all of Charlie's money.

Sam feels horrible.

Situation 11: Alex makes Chris wait.

Alex feels

Table 12: Prompt for generating xReact.

Next, what do people want in each situation? Examples:

Situation 1: Devin mows the lawn.

Devin wants to take a shower.

Situation 2: Jamie is going to a party.

Jamie wants to take an Uber home.

Situation 3: Sydney bleeds a lot.

Sydney wants to go to the ER.

Situation 4: Lindsay works as a cashier.

Lindsay wants to find a better job.

Situation 5: Rowan gets dirty.

Rowan wants to do a load of laundry.

Situation 6: Lee stays up all night studying.

Lee wants to rest.

Situation 7: Riley gets Noel's autograph.

Riley wants to tell some friends.

Situation 8: Adrian sees Taylor's point.

Adrian wants to agree with Taylor.

Situation 9: Hunter leaves Avery's bike.

Hunter wants to keep the bike safe.

Situation 10: Sam wants a tattoo.

Sam wants to find a tattoo design.

Situation 11: Alex makes Chris wait.

Alex wants

Table 13: Prompt for generating xWant.

Next, what can hinder each situation? Examples:

Situation 1: Devin makes a doctor's appointment,

This is hindered if Devin can't find the phone to call the doctor.

Situation 2: Jamie rubs Wyatt's forehead,

This is hindered if Jamie is afraid to touch Wyatt.

Situation 3: Sydney eats peanut butter,

This is hindered if Sydney is allergic to peanuts.

Situation 4: Lindsay looks perfect,

This is hindered if Lindsay can't find any makeup.

Situation 5: Rowan goes on a run,

This is hindered if Rowan injures her knees.

Situation 6: Lee takes Ali to the emergency room,

This is hindered if Ali has no health insurance to pay for medical care.

Situation 7: Riley spends time with Noel's family,

This is hindered if Noel's family doesn't like spending time with Riley.

Situation 8: Adrian moves from place to place,

This is hindered if Adrian can't afford to move.

Situation 9: Hunter protests the government,

This is hindered if Hunter is arrested.

Situation 10: Sam has a huge fight,

This is hindered if Sam does not like confrontation.

Situation 11: Alex makes Chris wait,

This is hindered if

Table 14: Prompt for generating HinderedBy.