

(COMET-)ATOMIC₂₀: On Symbolic and Neural Commonsense Knowledge Graphs

Jena D. Hwang^{†*}, Chandra Bhagavatula^{†*}, Ronan Le Bras[†], Jeff Da[†], Keisuke Sakaguchi[†],
Antoine Bosselut^{♠†} and Yejin Choi^{♡†}

[†] Allen Institute for AI, WA, USA

[♡] Paul G. Allen School of Computer Science & Engineering, WA, USA

[♠] Stanford University, CA, USA

Abstract

Recent years have brought about a renewed interest in commonsense representation and reasoning in the field of natural language understanding. The development of new commonsense knowledge graphs (CSKG) has been central to these advances as their diverse facts can be used and referenced by machine learning models for tackling new and challenging tasks. At the same time, there remain questions about the quality and coverage of these resources due to the massive scale required to comprehensively encompass general commonsense knowledge.

In this work, we posit that manually constructed CSKGs will never achieve the coverage necessary to be applicable in all situations encountered by NLP agents. Therefore, we propose a new evaluation framework for testing the utility of KGs based on how effectively implicit knowledge representations can be learned from them.

With this new goal, we propose ATOMIC₂₀, a new CSKG of general-purpose commonsense knowledge containing knowledge that is not readily available in pretrained language models. We evaluate its properties in comparison with other leading CSKGs, performing the first large-scale pairwise study of commonsense knowledge resources. Next, we show that ATOMIC₂₀ is better suited for training *knowledge models* that can generate accurate, representative knowledge for new, unseen entities and events. Finally, through human evaluation, we show that the few-shot performance of GPT-3 (175B parameters), while impressive, remains ~12 absolute points lower than a BART-based knowledge model trained on ATOMIC₂₀ despite using over 430x fewer parameters.

1 Introduction

Commonsense understanding and reasoning remain longstanding challenges in general artificial intelligence. However, large-scale language models have brought tremendous progress in the sub-field of natural language processing. Such large-scale language models (Radford et al. 2018; Devlin et al. 2019; Brown et al. 2020) trained on extreme-scale data have been shown to effectively adapt to diverse downstream tasks, achieving significant performance gains across

*The authors contributed equally to this work.

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹ATOMIC₂₀ and ATOMIC-2020 can be used interchangeably, but for brevity we use ATOMIC₂₀ in this paper.

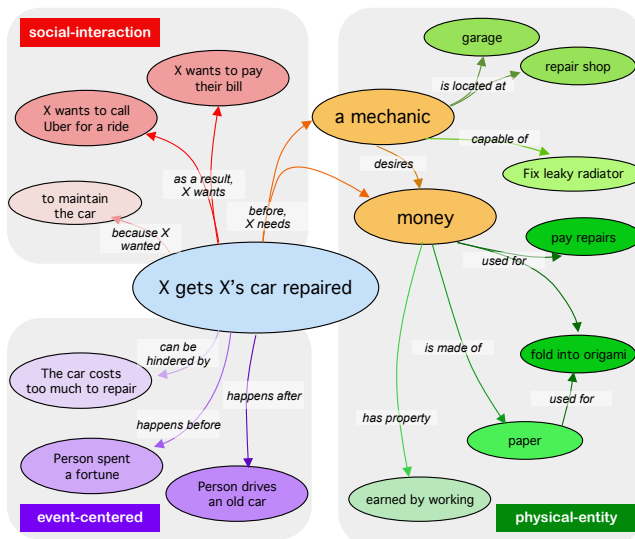


Figure 1: A tiny subset of ATOMIC₂₀, a large atlas of social and physical commonsense relations. Relations in the top-left quadrant reflects relations from ATOMIC.¹

natural language benchmarks (Wang et al. 2019). Interestingly, as these models have grown larger (and trained on larger amounts of data), their benchmark performance has continued to improve (Raffel et al. 2019) despite limited conceptual improvements, leaving open questions regarding the source of these remarkable generalization properties.

Recent work has hypothesized that many of these performance gains could be a result of language models being able to memorize facts in their parameters during training (Roberts, Raffel, and Shazeer 2020) that can be leveraged at evaluation time. As a result, a new paradigm of language models as knowledge bases has emerged (Petroni et al. 2019). In this setting, language models are prompted with natural language prefixes or questions, and they express knowledge through language generation. The initial success of this paradigm for representing commonsense knowledge (Davison, Feldman, and Rush 2019; Tamborrino et al. 2020) has led to the optimistic claim that language models comprehensively encode commonsense knowledge, and remove the need for structured knowledge resources.

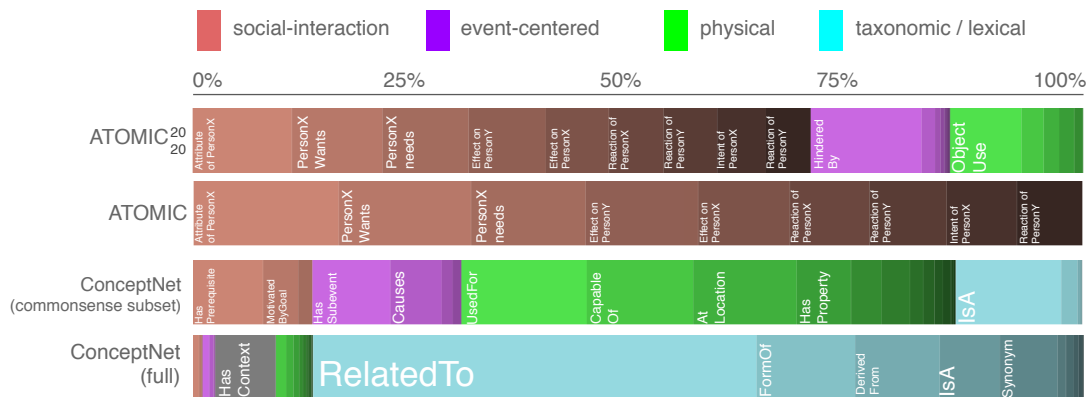


Figure 2: ATOMIC₂₀²⁰ tuple count distribution compared to ATOMIC (Sap et al. 2019) and CONCEPTNET, either its commonsense subset (Li et al. 2016) or the full set (Speer, Chin, and Havasi 2017).

We take a more skeptical view of this capacity of language models – *Does scaling up language models actually endow them with commonsense knowledge?* While language models can successfully express certain types of knowledge, their best results are observed in narrowly specific conditions – we show (cf. §5) that they perform better when evaluated on knowledge bases that prioritize ontological relations and whose examples resemble language-like assertions (e.g., mango `ISA` fruit).² Consequently, the types of knowledge that can be directly accessed through the language model’s interface remains limited.

However, prior work has also shown that training language models on knowledge graph tuples leads them to learn to express their implicit knowledge directly (Bosselut et al. 2019), allowing them to provide commonsense knowledge on-demand. These adapted *knowledge models* have exhibited promising results on commonsense benchmarks compared with methods that require linking entities to knowledge graphs (Shwartz et al. 2020; Liu et al. 2020). Inspired by these successes, we propose a dual use for commonsense knowledge bases going forward: as static graphs that can be linked to for discrete knowledge access, and as resources for adapting language models to hypothesize commonsense knowledge about un-annotated entities and events.

With this second purpose in mind, we propose evaluating commonsense knowledge resources based on the complementary information they can bring to pretrained language models. We construct ATOMIC₂₀²⁰, a new, high-quality knowledge graph with 1.33M commonsense knowledge tuples across 23 commonsense relations. We compare ATOMIC₂₀²⁰ with respect to its coverage and accuracy in competition with other highly used CSKGs, such as CONCEPTNET (Speer, Chin, and Havasi 2017). Our results show that ATOMIC₂₀²⁰ is able to cover more correct facts about more diverse types of commonsense knowledge than any exist-

ing, publicly-available commonsense knowledge resource. However, our results also indicate that there remains a large amount of exclusivity between these KGs, highlighting the challenge of creating resources that cover the scale and diversity of general commonsense knowledge.

Furthermore, we formalize the COMET framework of Bosselut et al. (2019) across different seed language models and training knowledge graphs, and evaluate the commonsense knowledge hypothesized by these adapted *knowledge models*. Our empirical study yields two promising conclusions. First, it confirms that KG-adapted language models learn to express knowledge more precisely than naive language models trained only on language. And second, we show that ATOMIC₂₀²⁰ as a transfer resource leads to COMET models that achieve the largest increase over their seed language model (across all seed LMs) for the commonsense knowledge types it covers, validating the importance of constructing knowledge resources with examples of knowledge not readily found in language models.

Key Contributions: In summary, we make three key contributions in this paper. We present ATOMIC₂₀²⁰—a new commonsense knowledge graph covering social, physical, and eventive aspects of everyday inferential knowledge (cf. §3). Next, we compare ATOMIC₂₀²⁰ with other prominent CSKBs head-to-head and show that our new *symbolic* knowledge graph is more accurate than any current CSKB (see Table 2) (cf. §4). Finally, we show that our new *neural* knowledge model COMET-ATOMIC₂₀²⁰ successfully transfers ATOMIC₂₀²⁰’s declarative knowledge to beat GPT-3, the largest pre-trained language model, in spite of using 400x fewer parameters (see Table 6) (cf. §5). This demonstrates the utility and importance of high-quality symbolic knowledge provided by ATOMIC₂₀²⁰ to generalize on commonsense information that LMs cannot expressively capture on their own (cf. §6).

2 Background

Commonsense Knowledge Graphs Large scale commonsense knowledge graphs are ubiquitous tools in natu-

²An observation supported by Brown et al. (2020)’s GPT-3 model, whose best few-shot performance on commonsense knowledge benchmarks comes on the PhysicalQA (Bisk et al. 2020) and HellaSwag (Zellers et al. 2019) datasets.

ral language processing tasks as access to their facts allows models to learn to reason over commonsense knowledge to make predictions (Lin et al. 2019; Feng et al. 2020). In this work, we evaluate three existing knowledge graphs, CONCEPTNET, ATOMIC, and TRANSOMCS on their coverage and precision relative to our new resource ATOMIC₂₀³.

The CONCEPTNET (v5.7) knowledge graph (Speer, Chin, and Havasi 2017) consists of 36 relations focusing mostly on taxonomic and lexical knowledge (e.g., RelatedTo, Synonym, IsA) and physical commonsense knowledge (e.g., MadeOf, PartOf). CONCEPTNET (v5.7) contains 3.4M entity-relation tuples (in English) collected by crowdsourcing and merged with existing knowledge databases from DBPedia, WordNet, Wiktionary, and OpenCyc. Since the knowledge are derived from human efforts, the accuracy of CONCEPTNET (v5.7) knowledge is fairly high, though the quality does vary depending on the sources of knowledge and relation types. However, as highlighted in (Davis and Marcus 2015; Sap et al. 2019), and shown in Figure 2, the coverage of CONCEPTNET (v5.7) is limited to mostly taxonomic, lexical, and object-centric physical commonsense knowledge. In fact, out of 3.4M tuples, 90% of them correspond to taxonomic (e.g., IsA) or lexical (e.g., Synonym, RelatedTo) knowledge, making the commonsense portion of CONCEPTNET (v5.7) relatively small.

The ATOMIC (Sap et al. 2019) knowledge graph consists of 880K of tuples across 9 relations that cover social commonsense knowledge (e.g, X gets X’s car repaired xIntent to maintain the car), including dynamic aspects of events such as causes and effects, if-then conditional statements, and mental states. The ATOMIC dataset is collected and validated completely through crowdsourcing.

The TRANSOMCS (Zhang et al. 2020a) knowledge graph consists of 18.48M tuples that were automatically converted from syntactic parses of sentences from various web sources including Wikipedia, Yelp, and Reddit. The set of relations used for the mapping is copied from CONCEPTNET. Although TRANSOMCS is much larger than other commonsense knowledge graphs, the precision of the extracted knowledge is significantly lower compared to other resources (cf. §4), and performs poorly as an adaptation resource relative to other KGs (cf. §5).

Language Models as Knowledge Bases Recent work hypothesizes that pretrained language models represent commonsense knowledge implicitly (Petroni et al. 2019; Roberts, Raffel, and Shazeer 2020). However, the results motivating these observations are often limited to narrowly scoped subsets of commonsense knowledge that primarily include taxonomic knowledge (e.g., mango IsA fruit) and that are often found explicitly stated in text. However, commonsense facts are often implied (Gordon and Van Durme 2013), and as will be seen in our studies (cf. §4), state of the art neural models struggle to express implicit commonsense knowledge that involves complex relationships.

³We were unable to include Cyc (Lenat 1995) in our study due to the discontinuation of its research license and the cost of the commercial license (over \$1M). CONCEPTNET includes a subset of Cyc – OpenCyc.

To overcome this limitation, Bosselut et al. (2019) take the best of both worlds between commonsense knowledge graphs and pretrained language models. The commonsense transformer, or COMET, adapts pretrained neural language models by training on example tuples from commonsense knowledge graphs. It takes a head/source phrase and a relation (e.g., take a nap Causes) and generates the tail/target phrase (e.g., have energy). Bosselut et al. (2019) show that COMET trained on the CONCEPTNET and ATOMIC knowledge graphs is able to adapt to generate novel (and valid) commonsense knowledge tuples.

Importantly, these neural *knowledge models* can produce commonsense knowledge on-demand for any head entity that can be expressed through language. This flexibility allows them to be used out-of-the-box, and they have been applied to new, previously unexplored tasks, such as sarcastic comment generation (Chakrabarty et al. 2020), therapy chatbots (Kearns et al. 2020), and automated story plot generation (Ammanabrolu et al. 2020). These contributions show that progress on knowledge models opens up new downstream applications that were challenging to model before.

3 ATOMIC₂₀

We present ATOMIC₂₀, a commonsense knowledge graph with 1.33M everyday inferential knowledge tuples about entities and events. ATOMIC₂₀ represents a large-scale commonsense repository of textual descriptions that encode both the social and the physical aspects of common human everyday experiences, collected with the aim of being complementary to commonsense knowledge encoded in current language models. ATOMIC₂₀ introduces 23 commonsense relations types. They can be broadly classified into three categorical types: 9 commonsense relations of social-interaction, 7 physical-entity commonsense relations, and 7 event-centered commonsense relations concerning situations surrounding a given event of interest. The full inventory of ATOMIC₂₀ relations is listed in Table 1.

In terms of physical and event-centered commonsense, by far, the two largest new relations in ATOMIC₂₀ are ObjectUse and HinderedBy. For ObjectUse, we focused on *affordances* of everyday objects such as “popcorn bucket” that may be used for “holding popcorn” or “storing things”. For HinderedBy, we explore the notion that many events in real world can be defeasible (Lascarides and Asher 1991) by collecting hindrances to goals that may be useful for tasks such as counterfactual reasoning. For example X’s desires to adopt a cat may be hindered by finding out that X is allergic to cats, which would necessitate X to adjust future actions accordingly (say, opt for hypoallergenic options like tortoises).

In the case of ObjectUse, we collected over 130K everyday object-use pairs by asking crowdworkers for necessary objects and their uses for each event in ATOMIC₂₀. For example, given “X eats popcorn” we elicited items such as “popcorn bucket” with their various expected uses. The number also reflects *atypical* usages gathered in a separate pass where workers were asked to provide creative or resourceful but *feasible* uses of the objects. Given “popcorn

	Head	Relation	Tail	Size
PHYSICAL-ENTITY	bread	ObjectUse	make french toast	165,590
		AtLocation*	basket; pantry	20,221
		MadeUpOf	dough; wheat	3,345
		HasProperty*	cooked; nice to eat	5,617
	baker	CapableOf*	coat cake with icing	7,968
		Desires*	quality ingredients	2,737
Not Desires*		bad yeast	2,838	
EVENT-CENTERED	X runs out of steam	IsAfter	X exercises in the gym	22,453
		HasSubEvent	become tired	12,845
		IsBefore	X hits the showers	23,208
		HinderedBy	drinks too much coffee	106,658
		Causes	takes a break	376
		xReason	did not eat breakfast	334
SOCIAL-INTERACTION	X watches --- anyway	isFilledBy	the game; the TV	33,266
	X runs out of steam	xNeed	do something tiring	128,955
		xAttr	old; lazy; lethargic	148,194
		xEffect	drinks some water	115,124
		xReact	tired	81,397
		xWant	to get some energy	135,360
	X votes for Y	xIntent	to give support	72,677
		oEffect	receives praise	80,166
		oReact	grateful; confident	67,236
		oWant	thank X; celebrate	94,548

Table 1: Relations in $ATOMIC_{20}^{20}$ along with illustrative examples and their respective size. Relations that reflect semantically identical categories to CONCEPTNET is marked with an asterisk (*).

bucket”, for instance, one might “wear it as a hat” for, say, a costume party. For *HinderedBy*, we crowdsourced over 100K tuples of hindrances to existing $ATOMIC_{20}^{20}$ events, asking the workers to provide situations or events that might pose as deterrence should the event be considered an achievable goal (see Appendix A for further details). For social-interaction commonsense, we primarily incorporated tuples from $ATOMIC$, but also crowdsourced an additional 34K tuples using the same approach as Sap et al. (2019).

$ATOMIC_{20}^{20}$ also pulls commonsense tuples from the English subset of CONCEPTNET(v5.7) (latest version available; Speer, Chin, and Havasi 2017).⁴ Of the 3.4M English tuples in CONCEPTNET(v5.7), a small subset of 172K tuples was selectively chosen to be integrated into $ATOMIC_{20}^{20}$. This subset represents data carefully identified to reflect commonsense information dealing with qualita-

⁴A CONCEPTNET(v5.7) fact is considered English if both the head and tail concepts are marked with ‘/en/’ in the edge id.

Knowledge Base	Accept	Reject	No Judgment
$ATOMIC_{20}^{20}$	91.3	6.5	2.2
$ATOMIC$	88.5	10.0	1.5
CONCEPTNET	88.6	7.5	3.9
TRANSOMCS	41.7	53.4	4.9

Table 2: **Accuracy** - Percentage (%) of tuples in the knowledge base evaluated by human crowdworkers as either always true or likely (Accept), farfetched/never or invalid (Reject), or unclear (No Judgment).

tive human experiences. Among the eliminated data are tuples with edge weight ≤ 0.5 , dictionary or etymologically based knowledge (e.g., synonyms/antonyms, inflections), lexical hyper/hyponymic lexical relationships such as *IsA* or *InstanceOf*, and relations based on lexical co-occurrence (e.g., *RelatedTo* or *LocatedNear*), which are easily recoverable from language models.⁵ After selective removal of these relations and a post-processing step to ensure the removal of deterministic information such as geographic facts (e.g., “shenzhen” *AtLocation* “china”), tuples from each CONCEPTNET were examined for further splits or joins to align with the existing structure of $ATOMIC_{20}^{20}$. A random 10% tuples from each selected relations were then put through crowdsourced validity testing (akin to the process described later in §4). Tuples that were directly incorporated without further edits passed with an acceptance rate of 93% or higher. A subset of relations (i.e., *CapableOf*, *MadeUpOf*, *HasProperty*) were put through additional crowdsourcing to weed out tuples that were either invalid or found to hold prejudiced descriptions of human entities. In the end, only 5 relations (marked with an asterisk in Table 1) retain the CONCEPTNET’s original meaning with a few relations that are cognates in $ATOMIC_{20}^{20}$ (more details in Appendix A).

4 Symbolic Knowledge Graph Comparison

In this work, we compare our new $ATOMIC_{20}^{20}$ knowledge graph to three other prominent CSKGs: $ATOMIC$ (Sap et al. 2019), CONCEPTNET⁶ (Li et al. 2016), and TRANSOMCS (Zhang et al. 2020a). We measure the accuracy of tuples in each KG and compare the coverage of each CSKG w.r.t. other CSKGs head-to-head.

Accuracy Assessment

In order to assess the accuracy of the knowledge represented, 3K random instances were extracted from each of the knowledge graphs for a crowdsourced evaluation of the tuples.

Human Evaluation Setup. The evaluation was carried out through crowdsourcing on the Amazon Mechanical Turk platform. Workers were presented with knowledge tuples in

⁵CONCEPTNET 5.7 defines weight as “the strength with which this edge expresses this assertion”. A pilot crowdsource assessment step found any tuple with weight ≤ 0.5 unreliable w.r.t. its validity.

⁶Hereafter, as we focus on CSKGs, by ConceptNet, we refer to the commonsense subset, unless specified otherwise.

ATOMIC ₂₀ ²⁰	ATOMIC	Relation	CN	T-OMCS
92.3		AtLocation*	89.4	34.3
93.9		CapableOf*	84.4	50.0
94.6		Causes	90.0	50.0
96.9		Desires*	96.3	48.2
93.9		HasProperty*	86.3	52.4
82.3		ObjUse/UsedFor	96.3	31.6
98.5		NotDesires*	96.3	
96.9		HasSubevent	88.1	57.7
		HasFirstSubevent	93.8	52.4
		HasLastSubevent	95.6	38.2
		HasPrerequisite	94.4	30.0
75.4		MadeUpOf/MadeOf	88.1	15.9
		PartOf	71.9	46.5
		HasA	77.5	43.5
96.9		HinderedBy		
96.2		isAfter		
95.4		isBefore		
96.2		isFilledBy		
		ReceiveAction	84.4	56.4
91.5	86.3	oEffect		
91.5	87.7	oReact		
88.5	89.5	oWant		
87.7	91.0	xAttr		
80.8	87.2	xEffect		
93.1	89.9	xIntent/MotivByGoal	84.4	27.1
87.7	85.1	xNeed		
90.8	91.3	xReact		
96.2		xReason		
82.3	88.4	xWant/CausesDesire	90.0	35.9

Table 3: KG accuracy values broken down by relation. Gray cells indicate statistically significant difference from ATOMIC₂₀²⁰ values. Dark gray cells signal instances where ATOMIC₂₀²⁰ values are significantly higher than its counterpart KB. Relational *cognates* have been grouped together and *exact matches* are asterisked (*) (cf. Table 1).

the form of (*head, relation, tail*) for annotation. To expedite the human assessment of the tuples, each *relation* (e.g., `xWant` or `AtLocation`) was translated into a human-friendly natural language form (e.g., “as a result, PersonX wants” and “located or found at/in/on”, respectively; cf. Appendix B). The workers were asked to rate the tuples along a 4-point Likert scale: *always/often* – the knowledge assertion presented is always or often true, *sometimes/likely* – it is sometimes or likely true, *farfetched/never* – it is false or farfetched at best, and *invalid* – it is invalid or makes no sense. Any tuples receiving the former two labels are ranked as **Accept** and latter two as **Reject**. The workers were also given a choice to opt out of assessment if the concepts were too unfamiliar for a fair evaluation (**No Judgment**). Each task (HIT) included 5 tuples of the same relation type, and each tuple was labeled by 3 workers. For the results, we take the majority vote among the 3 workers.

Results. ATOMIC₂₀²⁰ outperforms other KGs in crowd-sourced accuracy as shown in Table 2.⁷ ATOMIC ties with CONCEPTNET with reasonably high accuracy, while TRAN-

SOMCS lags behind others with far lower accuracy. We provide a per-relation breakdown of accuracies in Table 3.

Between ATOMIC₂₀²⁰ and ATOMIC, the variations in the assessed accuracies are not found to be statistically significant. Among the ATOMIC₂₀²⁰ and CONCEPTNET relations that represent *exact matches* (marked with * in Table 3), the differences are either not statistically significant or when they are, ATOMIC₂₀²⁰ improves upon the associated facts, reflecting that the preprocessing stages of CONCEPTNET integration were helpful in improving the quality of these relations (§3). Among *cognates* in ATOMIC₂₀²⁰ and CONCEPTNET relations, two sets of relations fare significantly worse in ATOMIC₂₀²⁰ than in CONCEPTNET. In the case of `ObjectUse/UsedFor`, this is likely due to the fact that ATOMIC₂₀²⁰’s `ObjectUse` includes atypical affordances (cf. §3). In an annotation setting where workers are asked to evaluate the truth or likelihood of an assertion rather than feasibility of use, a portion of the atypical usages are seen as ‘farfetched’ and thus, rejected. In the case of `MadeUpOf/MadeOf`, there may be some room for improvement for ATOMIC₂₀²⁰. Unlike the ATOMIC₂₀²⁰’s `HasSubEvent` label that successfully joins together CONCEPTNET’s `HAS(FIRST/LAST)SUBEVENT` labels for an improved accuracy, ATOMIC₂₀²⁰’s `MadeUpOf` union of `MadeOf`, `PartOf`, and a subset of `HasA`, did not seem to have resulted in improved quality. The rest of the ATOMIC₂₀²⁰ cognates see a significantly higher or similar accuracy in comparison to CONCEPTNET.

Coverage Assessment

We make a pairwise comparison between the CSKGs to assess their coverage with regards to the commonsense knowledge they contain. For a reliable head-to-head comparison, we map relations and tuples between various KGs.

Mapping Relations. Since ATOMIC₂₀²⁰ is built on existing ATOMIC relations, we primarily need to align relations between ATOMIC₂₀²⁰ and CONCEPTNET. We carefully align them based on the definitions for the labels as supplied by the two graphs, then the resulting alignment was verified by sampling at random approximately 20 instances per relation.

Mapping Tuples. In order to resolve syntactic differences in how the concepts are expressed in each of the KGs (e.g., ATOMIC’s “PersonX eats breakfast” vs. CONCEPTNET’s “eat breakfast”), we preprocess each of the head and tail concepts of each tuple in each KG in the following manner: (1) the concept is lowercased and stripped of extra spaces, punctuations, and stopwords; (2) any exact tuple duplicates within each KB removed, and (3) remaining content words are lemmatized according to their POS category. For ATOMIC and ATOMIC₂₀²⁰, an extra step is added to remove mentions of “PersonX”, “PersonY” and “PersonZ” if occurring at the beginning of a string, and to replace with ‘person’ if they occur elsewhere (e.g., “PersonX greets PersonY”).

Metrics. We use two metrics to evaluate the coverage of knowledge graphs. For each pair of CSKGs, we compute precision and recall with respect to a target KG. **Coverage** (moderate agreement; Fleiss 1971).

⁷Overall inter-rater agreement measured by Fleiss’ κ of 0.46

Source KB↓	Target KB→			
	ATOMIC	CN	T-OMCS	ATOMIC ₂₀ ²⁰
ATOMIC	-	0.1	0.0	100.0
CONCEPTNET	0.3	-	5.5	45.6
TRANSOMCS	0.0	0.4	-	0.3
ATOMIC ₂₀ ²⁰	60.2	9.3	1.4	-

Table 4: **Coverage Precision** - Average number of times (in %) a tuple in Source KB is found in Target KB.

Source KB↓	Target KB→			
	ATOMIC	CN	T-OMCS	ATOMIC ₂₀ ²⁰
ATOMIC	-	0.3	0.0	60.1
CONCEPTNET	0.1	-	0.3	8.9
TRANSOMCS	0.0	7.6	-	1.3
ATOMIC ₂₀ ²⁰	100.1 [†]	47.8	0.4	-

Table 5: **Coverage Recall** - Average number of times (in %) a tuple in Target KB is found in Source KB. [†]This value is greater than 100 because multiple tuples in ATOMIC₂₀²⁰ can map to the same tuple in ATOMIC.

precision assesses the proportion of tuples in the source KG that are correct according to tuples in the target KG. **Coverage recall** reflects the proportion of tuples in the target KB that the tuples in the source KB successfully recalled.

Results. Tables 4 and 5 show a pairwise coverage precision and recall assessment among the CSKGs. ATOMIC₂₀²⁰ shows the widest coverage: ATOMIC₂₀²⁰ is able to recall all of ATOMIC (as expected) and just under half of CONCEPTNET. There is very little overlap between ATOMIC and CONCEPTNET, which is unsurprising as all of ATOMIC knowledge is focused on social behaviors CONCEPTNET does not cover while CONCEPTNET leans on physical commonsense which falls outside ATOMIC’s scope. Overall, TRANSOMCS intersects very little with any of the other three KBs.

5 Neural Knowledge Graph Comparison

Language models are powerful tools for representing knowledge, but their ability to serve as generative knowledge bases is limited by the fact they are directly trained to represent the distribution of language. Previous work shows knowledge graphs can help language models better transfer as knowledge engines (Bosselut et al. 2019) by re-training them on examples of structured knowledge. As a result, a new purpose for knowledge graphs is to be useful in helping language models generalize to hypothesizing knowledge tuples.

Experimental Setup. To evaluate whether knowledge graphs can help language models effectively transfer to *knowledge models*, we train different pretrained language models on the knowledge graphs described in Section 4, which we describe below:

GPT2 (Radford et al. 2019) is a Transformer (Vaswani et al. 2017) based language model. In our experiments, we use the largest GPT2 model, GPT2-XL, that has 1.5B parameters.

KG	Model	Accept	Reject	No Judgm.
ATOMIC ₂₀ ²⁰	GPT2-XL	36.6	62.5	0.9
	GPT-3	73.0	24.6	2.5
	COMET(GPT2-XL)	72.5	26.6	0.9
	COMET(BART)	84.5	13.8	1.7
ATOMIC	GPT2-XL	38.3	61.2	0.4
	COMET(GPT2-XL)	64.1	34.7	1.2
	COMET(BART)	83.1	15.3	1.6
CONCEPTNET	GPT2-XL	50.3	42.1	7.7
	COMET(GPT2-XL)	74.5	19.0	6.4
	COMET(BART)	75.5	17.9	6.6
TRANSOMCS	GPT2-XL	28.7	53.5	17.8
	COMET(GPT2-XL)	26.9	60.9	12.2
	COMET(BART)	23.8	65.9	10.3

Table 6: Human evaluation of generation accuracy (%). Each model uses greedy decoding to generate the *tail* of 5K randomly-sampled test prefixes (*head, relation*) from each knowledge graph. GPT2-XL, GPT-3 and BART have 1.5B, 175B and 440M parameters, respectively.

We fine-tune GPT2-XL on each of our CSKGs to predict the tail of a tuple (e.g., wheat) given the head (e.g., bread) and a relation (e.g., MadeUpOf). The hyperparameter settings used for training are described in more detail in Appendix C. Additionally, we use GPT2-XL in a zero-shot setting as a baseline to measure the effect of transfer learning on knowledge graphs. For fair comparison, we convert each relation manually to an English language prompt expecting the tail of each tuple as output generated by the model.

BART (Lewis et al. 2020) is a Bidirectional and Autoregressive Transformer, an adaptation from BERT (Devlin et al. 2019) that is better suited for natural language generation (e.g., translation, summarization). Additional training details are provided in Appendix C.

GPT-3 (Brown et al. 2020) is an autoregressive language model that has 175B (over 100X more parameters than GPT2-XL) parameters and is trained on a corpus of web text. We use the GPT-3 API to *prime* the language model to generate the tail for a given prefix – (*head, relation*) pair. Thus, GPT-3 is evaluated in a few-shot setting. Additional details of our implementation are provided in Appendix C.

Evaluation Setup. To assess language-to-knowledge transfer capabilities, we evaluate how language models generalize to new, unseen entities, concepts, or events. We split each knowledge graph into training, validation, and test sets such that the *heads* of the knowledge tuples do not overlap between these sets. This adversarial split forces the language models to generalize the relationships they learn from training on the knowledge graphs to the entities learned during language pretraining. Also, to avoid overpopulating the validation and test sets with generic *heads* (e.g., “I”, “You”, “He”, “We”, and “They” collectively account for over 2.2M tuple heads in TRANSOMCS), we enforce that the head of any knowledge tuple in the *dev* and *test* sets is involved in at

		Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L	CIDeR	BERT Score
ATOMIC ₂₀	GPT2-XL	0.101	0.028	0.010	0.003	0.082	0.098	0.047	0.395
	GPT-3	0.299	0.153	0.081	0.048	0.182	0.255	0.175	0.540
	COMET(GPT2-XL)	0.407	0.248	0.171	0.124	0.292	0.485	0.653	0.638
	COMET(BART)	0.469	0.286	0.189	0.130	0.330	0.495	0.658	0.639
ATOMIC	GPT2-XL	0.083	0.029	0.011	0.005	0.081	0.087	0.045	0.386
	COMET(GPT2-XL)	0.419	0.296	0.228	0.189	0.292	0.517	0.733	0.634
	COMET(BART)	0.515	0.324	0.220	0.159	0.347	0.546	0.740	0.646
CONCEPTNET	GPT2-XL	0.044	0.012	0.004	0.002	0.064	0.057	0.050	0.389
	COMET(GPT2-XL)	0.155	0.119	0.095	0.078	0.134	0.193	0.425	0.552
	COMET(BART)	0.172	0.111	0.072	0.049	0.130	0.184	0.368	0.535
TRANSOMCS	GPT2-XL	0.028	0.001	0.000	0.000	0.093	0.053	0.013	0.351
	COMET(GPT2-XL)	0.301	0.000	0.000	0.000	0.180	0.302	0.254	0.677
	COMET(BART)	0.351	0.170	0.003	0.000	0.198	0.352	0.297	0.678

Table 7: Automated metrics for the quality of the *tail* generations of the GPT2-XL language model and the knowledge models COMET(GPT2-XL) and COMET(BART). Each approach uses greedy decoding for sampled 5k test prefixes for each KG. The 5k prefixes correspond to the ones for the human eval. Similar results are obtained on the full test sets (cf. Appendix C).

most 500 tuples. Finally, we remove low-quality tuples from TRANSOMCS by imposing a confidence score of ≥ 0.5 .

We score the tuples generated by these knowledge models using common evaluation metrics for text generation: BLEU (Papineni et al. 2002), ROUGE (Lin 2004), CIDeR (Vedantam, Lawrence Zitnick, and Parikh 2015), and BERT Score (Zhang et al. 2020b). For a subset of 5000 generated tuples from the test set of each knowledge graph, we also run the same human evaluation described in Section 4.

Results. We present our main results in Tables 6 and 7. First, we note the large divide between the zero-shot GPT2-XL model that produces commonsense knowledge without any fine-tuning and the two COMET models across the ATOMIC₂₀, ATOMIC, and CONCEPTNET knowledge graphs (Table 6). This large gap indicates that language models can benefit from learning facts from commonsense knowledge graphs. They do not have the means to precisely express this knowledge directly from just pretraining on language. This observation is supported by the gaps between these models in the automatic evaluations (Table 7), as well. Additionally, human evaluation of GPT-3 (Table 6) shows a ~ 12 point deficit compared to the performance of COMET(BART), in spite of GPT-3 (175B) having over ~ 430 times more parameters than COMET(BART) (406M). Similarly, we see a large gap in performance across all automated metrics in Table 7. The performance gap indicates that high-quality declarative knowledge is valuable even after the advent of extreme scale language models.

In addition to this main result, two particularly interesting observations emerge. First, we note that the gap between the zero-shot model and COMET is larger on the ATOMIC₂₀ and ATOMIC knowledge graphs, than on CONCEPTNET, supporting the reflection that ATOMIC₂₀ supports categories of knowledge that are more difficult to learn from pretraining. Second, the results on the human evaluation show that COMET models trained on TRANSOMCS are not able to generalize knowledge to new entities, implying

that language models benefit more from accurate knowledge examples, which TRANSOMCS lacks (cf. §4).

6 Discussion

Do pretrained language models already encode commonsense knowledge? Our conclusions on this subject are mixed and hinge on the ambiguous meaning of what it means to *encode* knowledge. Despite the conclusions of prior work (Petroni et al. 2019; Roberts, Raffel, and Shazeer 2020; Tamborrino et al. 2020), our results in Table 6 are clear that language models fail to express large varieties of knowledge when prompted for it in a zero-shot manner. When converted to COMET models by training on a knowledge graph, their performance at hypothesizing knowledge tuples skyrockets – 47.9% absolute difference between COMET(BART) and GPT2-XL on ATOMIC₂₀.

However, the evaluation tuples are adversarially selected to not include head entities that were in the training set. The model must generalize its learned representations of relations to entities it has not observed these relationships for during fine-tuning, meaning the representation of these entities is solely formulated from learning language. As a result, language models may still *encode* this knowledge in their parameters, even if they are not capable of *expressing* it directly. With this framing in mind, the COMET training paradigm proposed by Bosselut et al. (2019) can perhaps be viewed less as a means of learning *knowledge* from KGs, and more as a method of learning an *interface* for language models to hypothesize encoded knowledge through language generation. We look forward to future work in this space that attempts to disentangle these two ideas.

What considerations should be made when designing commonsense knowledge resources? Based on our results in Section 5, we outline desiderata for the design and development of future commonsense knowledge graphs. Because certain types of knowledge are already encoded and expressible by pretrained language models, CSKG designers should

focus on collecting examples and categories of knowledge that are less likely to be known by language models. For example, of the 378 test tuples evaluated by the GPT2-XL zero-shot model that contained the `HinderedBy` relation, only 1.3% were deemed plausible by human raters – jumping to 85% plausibility for COMET(BART) – pointing to an advantage in constructing `ATOMIC20` with this relationship in mind (see Appendix C for per-relation accuracy.).

Second, commonsense knowledge resources should be designed with the goal of accuracy and relationship coverage. Because language models exhibit powerful adaptation (Brown et al. 2020), they can generalize many commonsense relationships as long they have examples on which to train. Consequently, we should construct commonsense resources that encapsulate larger numbers of relations so the knowledge in pretrained language models can be grounded to a variety of relationships. However, language models also benefit from learning from precise examples. Being able to train on a large collection of examples from TRANSOMCS (see Appendix C) did not allow COMET models to generalize to unseen entities as these examples were not of sufficient quality (See Table 2). Resources should be carefully validated for the quality of their facts, an example set by Speer, Chin, and Havasi (2017) and Sap et al. (2019).

7 Conclusion

In this work, we formalize a use for commonsense knowledge graphs as transfer learning tools for pretrained language models. With this new purpose, we hypothesize that commonsense knowledge graphs should be designed to contain knowledge that is not already expressible by language models without difficulty (e.g., not taxonomic and lexical knowledge). Consequently, we propose `ATOMIC20`, a novel commonsense knowledge graph containing tuples whose relations are specifically selected to be challenging for pretrained language models to express. Our empirical studies demonstrate that `ATOMIC20` contains high-accuracy knowledge tuples across multiple novel relations not found in existing CSKGs or expressible by LMs. Furthermore, we show that `ATOMIC20` can be effectively used as a training set for adapting language models as *knowledge models* to generate high quality tuples on-demand.

References

Ammanabrolu, P.; Cheung, W.; Broniec, W.; and Riedl, M. 2020. Automated Storytelling via Causal, Commonsense Plot Ordering. *ArXiv abs/2009.00829*.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. G. 2007. DBpedia: A Nucleus for a Web of Open Data. In *ISWC/ASWC*.

Bisk, Y.; Zellers, R.; Le Bras, R.; Gao, J.; and Choi, Y. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *AAAI*.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Çelikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *ACL*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krüger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. *ArXiv abs/2005.14165*.

Chakrabarty, T.; Ghosh, D.; Muresan, S.; and Peng, N. 2020. R³: Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge. In *ACL*.

Davis, E.; and Marcus, G. 2015. Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence. *Commun. ACM* 58(9): 92–103.

Davison, J.; Feldman, J.; and Rush, A. 2019. Commonsense Knowledge Mining from Pretrained Models. In *EMNLP-IJCNLP*, 1173–1178. Hong Kong, China.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Feng, Y.; Chen, X.; Lin, B. Y.; Wang, P.; Yan, J.; and Ren, X. 2020. Scalable Multi-Hop Relational Reasoning for Knowledge-Aware Question Answering. *ArXiv abs/2005.00646*.

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

Gordon, J.; and Van Durme, B. 2013. Reporting bias and knowledge acquisition. In *AKBC '13*. ACM.

Herskovits, A. 2009. Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English.

Kearns, W. R.; Kaura, N.; Divina, M.; Vo, C. V.; Si, D.; Ward, T. M.; and Yuwen, W. 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*.

Landau, B.; and Jackendoff, R. 1991. Spatial language and spatial cognition. *Bridges Between Psychology and Linguistics: A Swarthmore Festschrift for Lila Gleitman* 145.

Lascarides, A.; and Asher, N. 1991. Discourse relations and defeasible knowledge. In *ACL*.

Lenat, D. B. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Compression. In *ACL*.

Li, X.; Taheri, A.; Tu, L.; and Gimpel, K. 2016. Commonsense Knowledge Base Completion. In *ACL*.

Lin, B. Y.; Chen, X.; Chen, J.; and Ren, X. 2019. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In *EMNLP/IJCNLP*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*.

- Liu, Y.; Yang, T.; You, Z.; Fan, W.; and Yu, P. S. 2020. Commonsense Evidence Generation and Injection in Reading Comprehension. In *SIGDIAL*.
- Miller, G. A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*.
- Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A. H.; and Riedel, S. 2019. Language Models as Knowledge Bases? In *EMNLP*.
- Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training.
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.
- Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv abs/1910.10683*.
- Roberts, A.; Raffel, C.; and Shazeer, N. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?
- Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lorie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *AAAI*.
- Shwartz, V.; West, P.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2020. Unsupervised Commonsense Question Answering with Self-Talk. *ArXiv abs/2004.05483*.
- Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.
- Talmy, L. 1988. Force Dynamics in Language and Cognition. *Cognitive Science* 12: 49–100.
- Tamborrino, A.; Pellicanò, N.; Pannier, B.; Voitot, P.; and Naudin, L. 2020. Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning. In *ACL*.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.
- Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, 4566–4575.
- Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *ArXiv abs/1905.00537*.
- Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2019. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *ArXiv abs/1910.03771*.
- Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *ACL*.
- Zhang, H.; Khashabi, D.; Song, Y.; and Roth, D. 2020a. TransOMCS: From Linguistic Graphs to Commonsense Knowledge. In *IJCAI*.
- Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.; and Artzi, Y. 2020b. BERTScore: Evaluating Text Generation with BERT. *ArXiv abs/1904.09675*.

A ATOMIC 2020 Details

ATOMIC₂₀²⁰ Relations

In this section we detail the relations in ATOMIC₂₀²⁰. Figure 3 shows the hierarchical breakdown of the ATOMIC₂₀⁵⁰ relation labels. While there is no internal structure directly encoded for ATOMIC₂₀²⁰ relations, they fall into three natural categories based on their meaning: physical-entity, social-interaction and event-centered commonsense.

Physical-Entity Commonsense. Physical-entity commonsense deals with inferential knowledge about common entities and objects. Physical commonsense such as these is crucial for interacting with the world: allowing us to distinguish the dangerous (e.g., “fire can be painful”) from the harmless (e.g., “teddy bears are comforting”), manipulate objects for our use (e.g., “helmets protect head”), and solve problems (e.g., “how do I open this door?”). We identify **seven** relations under this category.

- **ObjectUse** describes everyday **affordances** or uses of objects, and includes both *typical* and *atypical* uses. For example, “popcorn bucket” can typically be used to “hold popcorn” but it could also serve “as a hat” in atypical situations.
- **MadeUpOf** and **HasProperty**, two **property** relations, denote the relationship between an entity and its composition or characteristics. **MadeUpOf** describes a part, portion or makeup of an entity. For example, “cake” can be **MadeUpOf** “eggs” (composition/ingredient) or “icing” (part/portion). Similarly, **HasProperty** usually describes entities’ general characteristics such as “rose” is “red,” subjective attributes such as “thirst” is “uncomfortable.” In certain cases, the relation can also map to descriptors that speak to the substance or value of items such as “meat” has property of being “stored in the freezer” or “bike” is “powered by person’s legs.”
- **AtLocation** is a **spatial** relation that describes the location in/on/at which an entity is likely to be found (e.g. “gambler” can be found in “casino,” “wrench” can be found in “garage”).
- **CapableOf** is designed to describe abilities and **capabilities** of everyday living entities (e.g., humans, animals, insects) and natural entities that can exert a force (e.g. sun, storms). **CapableOf** includes general capabilities such as a “human” is capable of “thinking and reasoning” or “drinking coffee.” It also includes specialized capabilities such as a “surgeon” is capable of “operating on a patient.”
- **Desires** and **NotDesires** are relations that deal with **desires**⁸ of sentient entities; e.g., “doctors” likely desire to “cure patient” but do not desire “malpractice suit.”

⁸Since desire relations are about cognitive states of sentient beings, they also provide a degree of commonsense about social-interaction. However, we point out that these relations indicate *generic* characterizations of animate entities rather than describing situationally-based cognitive mental states (e.g., X being ‘encouraged’ only applies to the event it is situated in). For this reason, we include these relations under physical-entity commonsense.

Social-Interaction Commonsense. Social-interaction relations comment on socially-triggered states and behaviors. Social commonsense is useful for gauging people’s intentions and purpose, and predicting situationally-relevant human reactions and behaviors. Following the definitions for ATOMIC relations (Sap et al. 2019), we identify a total of **nine** relations within this category.

- Three **mental state** relations address the emotional or cognitive states of the participants in a given event. **xIntent** defines the likely *intent* or desire of an agent (X) behind the execution of an event. Given the head “X gives Y gifts,” an **xIntent** might be that X wanted “to be thoughtful.” Relations **xReact** and **oReact** define the *emotional reactions* on the part of X or other participants in an event. As a result of gift giving, X might feel “good about [one]self” and others (in this case, Y) might feel “appreciated.”
- Five **behavioral** relations address the socially relevant responses to an event. **xNeed** describes a precondition for X achieving the event. For example, in order for X to give Y gifts, X must first “buy the presents.” **xWant** and **oWant** are postcondition desires on the part of X and others, respectively. As a result of X giving Y gifts, X may also desire “to hug [Y]” and Y may want to “open the gift.” **xEffect** and **oEffect** are social actions that may occur after the event: X may “get hugged” and Y may “blush” in response.
- The last relation **xAttr** describes X’s **persona** or attribute as perceived by others given an event. In the gift giving example, X may be seen as “generous” or “giving.” In contrast, in an event such as “X steals a car,” X may be perceived as “evil.”

Event-Centered Commonsense. While social-interaction commonsense gauges human behaviors and mental states given an event, the event-centered commonsense provides intuitions about how common events are related to one another. Commonsense about event interaction is useful for understanding likely causes and effects of events in the world. This knowledge allows humans to strategize and explore the best solutions for their objectives, make contingency plans, and revise goals when circumstances deviate from expectation. There are **seven** relations that fall under this category.

- We group three relations under **force dynamics**.⁹ This group conceptualizes dynamic interactions between events with regards to exerted causal forces and impelled actions. **Causes** specifically captures the causal relation between two events or entities – e.g. an “accident” can cause “injury.” **Causes** does have some overlap with behavioral relations such as **xEffect** in that they are postconditions of an event, but the postcondition in **Causes** is not socially triggered and can exist outside human control (e.g., “bad weather” causes “power

⁹For a discussion of force dynamics in cognitive linguistic and lexical semantic literature cf. Herskovits (2009); Landau and Jackendoff (1991); Talmy (1988).

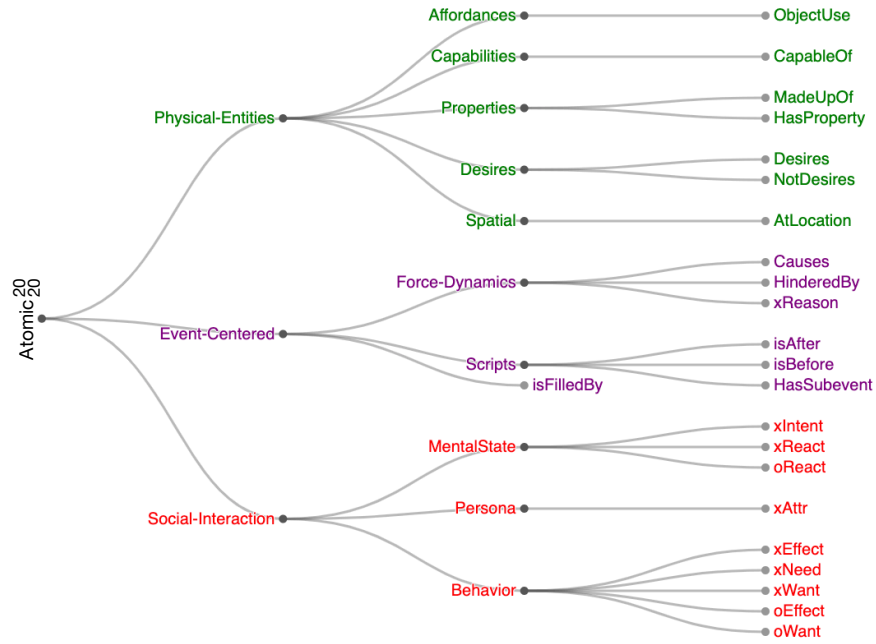


Figure 3: ATOMIC₂₀ relations organized into a hierarchical structure.

outages”). **HinderedBy** introduces hindrances that obstruct the natural path to the achievement of a goal. For example, the event “X adopts a cat” can be obstructed if “X is allergic to cats.” **xReason** provides a post-fact explanation of the cause of an event (e.g., why one has to “walk” could be explained by “car has broken down”), which is related to, but distinct from, **xIntent**’s intentions (i.e., “X walks” because X wanted to “go home”).

- Three relations provide reasoning about event **scripts** or sequences. **isAfter** and **isBefore** introduce events that can precede or follow an event, respectively. For example, “X is in a hurry to get to work” can happen after “X wakes up 15 minutes late” and before “X drives too fast.” These relations are distinguished from behavioral relations **xNeed** (pre-condition) and **xEffect** (post-condition) in that **isAfter** and **isBefore** are temporally situated without specific regard to the need or reaction of the person X. For example, “X pushes X’s luck” can happen before “X gets a broken nose” but getting a broken nose is not an action X intentionally may take after pushing one’s luck. Relation **HasSubEvent** provides the internal structure of an event, each tail denoting a step within the larger head event.
- The last relation in the event-centered category, **isFilledBy**, provides a filler phrase for an event with a blank that is sensical and commonly acceptable for the event. For example, the blank in an event such as “X catches ___ in the act” can be commonly filled by entities such as a “cheater,” a “burglar,” or a “mouse.”

ATOMIC₂₀ Tuples

In this section, we detail the population of the ATOMIC₂₀ tuples (see Table 1 for counts per relation).

Social-Interaction Tuples. For social-interaction relations, we incorporated 877K tuples from ATOMIC, and crowdsourced an additional 34K tuples using the same approach as Sap et al. (2019). The rest of this section will refer to the head events in the social-interaction tuples as **base events**.

Crowdsourced Tuples. Tuples for relations **ObjectUse**, **HinderedBy**, **isFilledBy**, **isBefore** and **isAfter** were crowdsourced via Amazon Mechanical Turk. We paid an average of \$15 an hour for our crowdsourcing efforts. We release all crowdsourcing templates as part of our code-base.¹⁰

- For the collection of **HinderedBy**, we crowdsourced over 100K event-hindrance tuples by prompting the workers with base events from ATOMIC and eliciting reasons why one may not be able to achieve the event. In order to make the prompt events more worker-friendly, we processed the events as a desire (e.g., “X adopts a cat” → “X wants to adopt a cat”).¹¹ We specifically elicited personal causes (e.g., “X is allergic to cats”), situational causes (e.g., “there are no pet stores nearby”), and social causes (e.g., “X’s landlord disallows pets”).
- 33K **isFilledBy** tuples were collected by presenting workers with base events. The workers were asked to provide two (up to four) common objects or entities that will make sense in the sentence.
- 46K tuples for **isBefore** and **isAfter** were collected together as sequences of events. Given a base event, the workers were asked to write a short 3-sentence story by

¹⁰<http://anonymous>

¹¹To achieve this, we removed modal verbs, lemmatized the head verb of the sentence, and inserted a ‘want to’ phrase before the verb.

CONCEPTNET	ATOMIC ₂₀ ²⁰
AtLocation	AtLocation
CapableOf	CapableOf
Causes	Causes , xEffect
CausesDesire	xWant
Desires	Desires
MadeOf	MadeUpOf
HasProperty	HasProperty
HasA	MadeUpOf , HasProperty
HasPrerequisite	xNeed
HasSubevent	HasSubEvent
HasFirstSubevent	HasSubEvent
HasLastSubevent	HasSubEvent
NotDesires	NotDesires
MotivatedByGoal	xIntent , xReason
PartOf	MadeUpOf
UsedFor	ObjectUse
ReceivesAction	MadeUpOf , AtLocation, Causes, ObjectUse

Table 8: CONCEPTNET relations mapped to ATOMIC₂₀²⁰ relations. For labels mapping to multiple ATOMIC₂₀²⁰ relations, the one that received the majority mapping is bolded.

providing a preceding and following event. The workers were given the option to opt out of writing a story if they felt that the event they were given didn't make sense enough to create a story.

- As discussed in the main text (§3), 130K ObjectUse tuples were crowdsourced by eliciting common objects and their uses for every event in the collected event sequences. For each event, a worker was asked to provide 2-4 common items that were needed during the displayed event. Atypical ObjectUse was collected in a second pass, where for each collected unique object, the workers were prompted with the object and asked to provide an atypical, creative or resourceful use for the item shown.

Integration of CONCEPTNET Tuples. The tuples for the remaining relations are populated through the integration of the commonsense portion of CONCEPTNET. As discussed in the main text, a select subset of CONCEPTNET(v5.7) tuples (172K) were integrated into ATOMIC₂₀²⁰.

The primary challenge in integrating CONCEPTNET tuples into ATOMIC₂₀²⁰ was in identifying knowledge that is most likely to reflect commonsense information. CONCEPTNET(v5.7) contains tuples built on not only concept relationships directly sourced from human informants, but also on information pulled from other lexical sources such as WordNet (Miller 1995) and DBpedia (Auer et al. 2007), which automatically extracts knowledge from Wikipedia articles (Speer, Chin, and Havasi 2017). As a result, even those relations that are *designed* to primarily represent commonsense knowledge (i.e., the OMCS relations) include among the mix, tuples that reflect factual or lexical co-occurrence knowledge. These examples deviate from the type of knowledge we would ideally consider as “commonsense,” i.e., qualitative experiential knowledge gained through subjective observation of and interaction with the world. Relations

Relations	Human Readable Template
AtLocation	located or found at/in/on
CapableOf	is/are capable of
Causes	causes
CausesDesire	makes someone want
CreatedBy	is created by
Desires	desires
HasA	has, possesses or contains
HasFirstSubevent	BEGINS with the event/action
HasLastSubevent	ENDS with the event/action
HasPrerequisite	to do this, one requires
HasProperty	can be characterized by being/having
HasSubEvent	includes the event/action
HinderedBy	can be hindered by
InstanceOf	is an example/instance of
isAfter	happens after
isBefore	happens before
isFilledBy	blank can be filled by
MadeOf	is made of
MadeUpOf	made (up) of
MotivatedByGoal	is a step towards accomplishing the goal
NotDesires	do(es) NOT desire
ObjectUse, UsedFor	used for
oEffect	as a result, Y or others will
oReact	as a result, Y or others feels
oWant	as a result, Y or others want
PartOf	is a part of
ReceivesAction	can receive or be affected by the action
xAttr	X is seen as
xEffect	as a result, PersonX will
xIntent	because PersonX wanted
xNeed	but before, PersonX needed
xReact	as a result, PersonX feels
xReason	because
xWant	as a result, PersonX wants

Table 9: Human readable templates for each relation used for crowdsourced human evaluations.

such as InstanceOf (“is instance/example of”) stands as a case in point (e.g., “tortilla” is an example of “flatbread” or “toffee” is an example of “candy”). While included within the OMCS relations, the encoded information can be hard to distinguish from the more accepted taxonomic relations such as IsA (“is a kind/type of”).¹² Relationships found in relations such as RelatedTo and DistinctFrom are too underspecified with regards to the meaning they represent, and for other relations such as LocatedNear, and negative forms such as NotCapableOf or NotHasProperty, the relationships amount to general lexical relationships.

Thus, the process of CONCEPTNET(v5.7) knowledge selection (described in §3) was judiciously guided by three competing priorities: when possible, we prioritized (1) qualitative commonsense over factual knowledge, (2) general knowledge over highly specific knowledge (e.g., personal names), and (3) meanings that are specific enough to be

¹²In fact, CONCEPTNET(v5.7) recognizes the similarities between IsA and InstanceOf and has accordingly deprecated InstanceOf in favor of IsA. Nevertheless, InstanceOf is still found in CONCEPTNET(v5.7).

meaningfully categorized. Since the ideal route of verifying imported data via crowdsourcing can be resource-intensive, we opted for an approach whereby relations were first selected based on the data they represent; then tuples were pruned based on heuristics that leverage lexical and syntactic information of the concepts. As mentioned in the main text, 10% of the data selected for integration was validated by crowdworkers, yielding a greater than 93% acceptance rate. Three relations, namely `HasProperty`, `CapableOf`, and `MotivatedByGoal`, were sent for instance-by-instance crowdsourcing for the purpose of debiasing human-related descriptions, and subdividing semantically distinct elements within the category (e.g., `MotivatedByGoal` mapped to `xIntent` and `xReason`). The resulting CONCEPTNET-to-ATOMIC₂₀ relation mapping details are shown in Table 8.

B Symbolic Knowledge Graph Details

Human Evaluation

Qualifying Crowdsource Workers. To ensure high-quality annotations, we qualified a pool of 173 workers through a paid qualification task that tested their ability to follow directions and provide reasonable answers to the qualification test. The qualification test contained 6 manually selected tuples from ATOMIC and CONCEPTNET, including both easy and tricky relations to annotate. A worker was qualified if they provided 100% acceptable answers. Workers providing 5 of 6 correct answers were also accepted only when they provided a reasonable written substantiation for their incorrect choice. Workers were paid an average of \$15 per hour for their evaluations.

Human Readable Relation Templates. Since the KB relation labels are rather telegraphic on their own, we used human readable language forms (based ATOMIC₂₀ and CONCEPTNET definitions) for prompt display in crowdsourced evaluations. The complete list is available in Table 9.

KB Accuracy & Coverage

In Table 2, what type of tuples generally end up with no judgment? Tuples receiving *no judgment* fall into three general categories: (1) either the head or the tail concept is too specialized for the workers to judge without consulting reference (e.g., “klebsiella” is part of “bacteria,” “drug cocktail” made of “nucleoside reverse transcriptase inhibitor”); (2) concepts refer to highly specific entities or referents (e.g., “singh” capable of “bring key,” “falkland island islas malvinas” part of “argentina”); and (3) *Reject* candidates that workers have decided to hedge on (e.g., “dandelion” used for “love,” “democrat” desires “matter”). Such tuples are mostly found in TRANSOMCS, as evidenced by the high fraction of tuples that received *No Judgment* at less than half of ATOMIC₂₀’s *Accept* rate (see 2).

Does the accuracy ratings breakdown for each KB provide further insights? A closer look at the raw accuracy ratings shows an interesting emergent rating pattern across KBs (Table 2). For all KBs with the exception of TRANSOMCS, we observe that the majority of social-interaction *Accept* originate from the *sometimes/likely* rating. However, such

preference is not seen in the physical-entity tuples, which show a slightly higher tendency for the *always/often* rating. For event-centered tuples, ATOMIC₂₀ favors the *sometimes/likely*, while CONCEPTNET does not. TRANSOMCS shows highest ratings for the *sometimes/likely* and *invalid* ratings, and the patterns are invariant across the board.

One additional point to mention is that ATOMIC₂₀ social-interaction and event-centered tuples proportionally contain more of the human-crowdsourced commonsense knowledge than the physical-entity category, which, with the sole exception of `ObjectUse`, includes tuples integrated from CONCEPTNET graph. The observation that much of the knowledge in ATOMIC₂₀ is sometimes or likely true, reflects our intentional efforts to deprioritize factual information over qualitative commonsense knowledge. More importantly, it shows that most of the knowledge within the ATOMIC₂₀ graph can be, under the right circumstances, defensible. That is, one can pose a likely hypothesis that a hindrance to “X writes stories” is that “X can’t read;” however, such a hypothesis can be defeated if we also know that X has written stories before. We find that such context-dependent ambiguities are of more compelling interest to us, as certainties may be better covered by language models.

C Neural Knowledge Graph Details

Dataset Split. Table 10 reports the number of tuples for each three-way split (train/dev/test) of each knowledge graph. The ATOMIC₂₀ split preserves the splits from ATOMIC and CONCEPTNET: any tuple in ATOMIC₂₀ that appears in the train (resp. dev, test) set of ATOMIC or CONCEPTNET belongs to the train (resp. dev, test) set of ATOMIC₂₀. Overall, ATOMIC₂₀ provides over 50% more tuples than the initial version ATOMIC.

Knowledge Graph	Train	Dev	Test	All
ATOMIC ₂₀	1,076,880	102,024	152,209	1,331,113
ATOMIC	709,993	79,599	87,480	877,072
CONCEPTNET	264,791	5,000	30,209	300,000
TRANSOMCS	5,424,478	10,243	100,033	5,534,754

Table 10: Number of tuples per KB and per split.

Details about GPT2-XL Training. GPT2-XL (Radford et al. 2019) is a transformer language model trained on 8 million webpages (~40G of text data). We finetune the language model on each commonsense knowledge graph by converting a tuple into a formatted text input – e.g. `<head> <relation> [GEN] <tail> [SEP]`. where [GEN] and [SEP] are special delimiter tokens that indicate the start and end of the tail of a given relation for a given head entity / event. At inference time, the head and relation of a tuple are given as input and the model’s generation following the [GEN] token is recorded as its prediction of the tail entity. We finetuned GPT2-XL on each CSKG for one epoch, using a batch size of 32 and a learning rate of $5e - 5$ on an Nvidia RTX-8000 GPU. The final trained models for each CSKG will be publicly released as part of our code release. In Figure 5, we include a few examples of COMET(GPT2-XL)

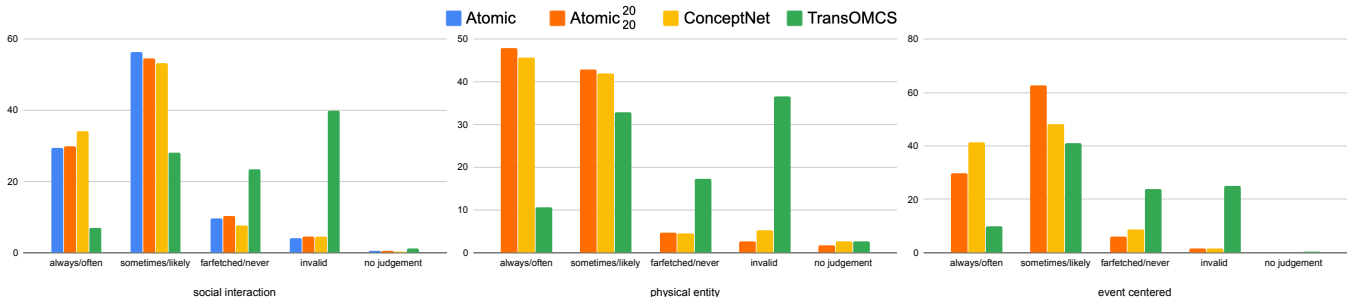


Figure 4: Percentage distribution of raw accuracy ratings broken down by KB (i.e., breakdown of Table 2). From left to right are the ratings for social-interaction tuples, physical-entity tuples, and event-centered tuples. We use the CONCEPTNET-to-ATOMIC²⁰ relation mappings (shown in Table 8) to categorize CONCEPTNET and TRANSOMCS relations into the three categories. For multiple mappings, we map the CONCEPTNET/TRANSOMCS labels to the majority mapped label (in bold in Table 8). Note that the latter two figures do not include ATOMIC as the KB only includes social-interaction relations.

		Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L	CIDEr	BERT Score
ATOMIC ²⁰	COMET(GPT2-XL)	0.401	0.247	0.168	0.123	0.288	0.473	0.620	0.632
	COMET(BART)	0.462	0.280	0.182	0.124	0.325	0.486	0.632	0.636
ATOMIC	COMET(GPT2-XL)	0.429	0.300	0.225	0.187	0.297	0.527	0.754	0.638
	COMET(BART)	0.521	0.330	0.225	0.164	0.351	0.552	0.766	0.650
CONCEPTNET	COMET(GPT2-XL)	0.152	0.115	0.092	0.080	0.131	0.193	0.421	0.552
	COMET(BART)	0.169	0.108	0.069	0.046	0.127	0.180	0.350	0.532
TRANSOMCS	COMET(GPT2-XL)	0.298	0.000	0.000	0.000	0.179	0.300	0.249	0.677
	COMET(BART)	0.351	0.216	0.004	0.000	0.201	0.352	0.298	0.681

Table 11: Automated metrics for the quality of the *tail* generations for the knowledge models COMET(GPT2-XL) and COMET(BART). Each approach uses greedy decoding for all test prefixes for each KG. Similar results were obtained on the 5K sampled prefixes that were randomly selected for the human evaluation (see Table 7).

trained on ATOMIC²⁰.

Details about BART Training. BART (Lewis et al. 2020) is a denoising sequence-to-sequence pretrained language model. Similar to previous transformer-based language models (Devlin et al. 2019), BART’s pretraining objective is to recover its input, which is corrupted through various strategies such as token and span masking, and sentence permutation. For pretraining, BART uses a 160GB free-text dataset drawn from news, books, stories, and web texts. We used the BART-large version of the model,¹³ which has 24 layers, 1024-dimensional hidden states, 16 attention heads in its self-attention layers, and 406M total parameters. For hyper-parameter search, we fine-tuned BART on each commonsense knowledge graph for one epoch with batch sizes {64, 32, 16}, learning rates {1e-3, 1e-5, 1e-7}, and three random seeds.

Details about GPT-3 Evaluation. We evaluate GPT-3 (Brown et al. 2020) using OpenAI’s language completion API. Similar to zero-shot evaluation on GPT2-XL, we use templates to evaluate the ability of the language model to generate a tail given the head and relation. We use the same templates as GPT2-XL. For priming examples, we prime each relation with 5 examples of heads and tails per relation,

randomly selected from relations in the training set. We ran 3 random seeds to select priming examples to avoid spelling mistakes and other fragments from data collection. We ran with temperature 0.4.

Additional Automated Evaluation. In order to have a direct comparison between automated and human evaluations, we report in Section 5 the automated metrics on the same test subsets that were used for human evaluation. For completeness, in this section, we provide the automated evaluation results on the full test sets (Table 11). These results confirm the findings of Section 5.

¹³from HuggingFace’s implementation (Wolf et al. 2019).

Head	Relation	Human	COMET (GPT2 XL)	COMET (BART)
PersonX accepts PersonY's apology	xIntent (Intent of PersonX)	To feel peaceful with PersonY	To be kind	To be forgiving
PersonX accepts PersonY's apology	HinderedBy	PersonY has not apologized	PersonX is stubborn	PersonX is too angry
PersonX accepts PersonY's apology	oReact (Reaction of PersonY)	Happy	Relieved	Happy
PersonX does yard work	isAfter	PersonX is off today	PersonX gets home from work	PersonX gets a job as a gardener
PersonX does yard work	isBefore	PersonX is proud of being productive	PersonX gets paid	PersonX takes a shower
PersonX abuses PersonX's power	xAttr (attribute of PersonX)	Out of line	Mean	Mean
PersonX abuses PersonX's power	xWant (Want of personX)	Harass employees	To get their way	To be in control
Common sense	ObjectUse	Prevent mistakes	Understand the situation	Make a good decision
Move car	HasSubEvent	Open car door	Turn steering wheel	Get out of car
PersonX loses PersonY's wallet	xWant (Want of PersonX)	To get the wallet	To apologize	to find the wallet
PersonX breaks the news to PersonY	oWant (Want of PersonY)	to cry	to cry	to cry
family photograph	AtLocation	relatives house	Photograph book	Family room
mouse	HasProperty	Whiskers	Four legs	Long tail
bird lover	CapableOf	listen to mockingbird sing	watch birds fly	Look at birds

Figure 5: Example generations of models on relations from ATOMIC₂₀²⁰. Red, purple and green rows represent social-interaction commonsense, event-centered commonsense, and physical-entity commonsense, respectively.

D Additional Reproducibility Items

All experiments were conducted on a cluster with 8 GPUs of type NVIDIA Quadro RTX 8000 with 48 GB of GDDR6 memory each. To allow replication of results, whenever possible, a default, fixed value was assigned to the random seed that initializes the pseudo-random number generator, as specified in the source code. The details of the experimentation of the models (i.e. GPT2-XL and BART), including their hyper-parameter settings, are described in Appendix C. All the data as well as the source code required for conducting experiments will be made publicly available upon publication.

