Unsupervised Commonsense Question Answering with Self-Talk

Vered Shwartz[1,2], **Peter West**[1,2], **Ronan Le Bras**[1], **Chandra Bhagavatula**[1], **and Yejin Choi**[1,2]
[1]Allen Institute for Artificial Intelligence
[2]Paul G. Allen School of Computer Science & Engineering, University of Washington
{vereds,peterw,ronanlb,chandrab,yejinc}@allenai.org

Abstract

Natural language understanding involves reading between the lines with implicit background knowledge. Current systems either rely on pre-trained language models as the sole implicit source of world knowledge, or resort to external knowledge bases (KBs) to incorporate additional relevant knowledge. We propose an unsupervised framework based on *self-talk* as a novel alternative to multiple-choice commonsense tasks. Inspired by inquiry-based discovery learning (Bruner, 1961), our approach inquires language models with a number of information seeking questions such as *"what is the definition of ..."* to discover additional background knowledge. Empirical results demonstrate that the self-talk procedure substantially improves the performance of zero-shot language model baselines on four out of six commonsense benchmarks, and competes with models that obtain knowledge from external KBs. While our approach improves performance on several benchmarks, the self-talk induced knowledge even when leading to correct answers is not always seen as helpful by human judges, raising interesting questions about the inner-workings of pre-trained language models for commonsense reasoning.

1 Introduction

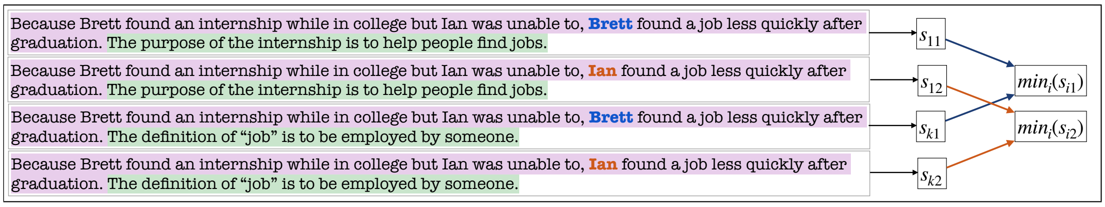
Human level natural language understanding involves reading between the lines and relying on implicit background knowledge. Consider the sentence: *Alice let Bob stand in front of her at the concert*. Using physical and social commonsense – (i) Bob and Alice want to see the stage, and (ii) If Bob is taller, they would block Alice's view – one can infer that Alice is taller than Bob. Such examples are ubiquitous across natural language understanding (NLU) tasks such as reading comprehension (Hirschman et al., 1999) and recognizing textual entailment (Dagan et al., 2013), and even more so in tasks dedicated to commonsense reasoning such as the Winograd schema challenge (Levesque et al., 2012). Most current NLU models rely on pre-trained language models (LMs; e.g. Radford et al., 2019; Devlin et al., 2019; Raffel et al., 2020). The standard practice is to fine-tune a pre-trained LM in a supervised manner on task-specific data. Alternatively, LM score is used to rank answer choices in a zero-shot setup (Wang et al., 2019; Bosselut and Choi, 2019). In both setups, pre-trained LMs yield improved performance upon prior methods, greatly due to the world knowledge that such LMs capture, having been trained on massive texts (Petroni et al., 2019; Davison et al., 2019).

Despite the performance boost, LMs as knowledge providers suffer from various shortcomings: (i) *insufficient coverage*: due to reporting bias, many trivial facts might not be captured by LMs because they are rarely written about (Gordon and Van Durme, 2013). (ii) *insufficient precision*: the distributional training objective increases the probability of non-facts that are semantically similar to true facts, as in negation ("birds cannot fly"; Kassner and Schütze, 2020). LMs excel in predicting the semantic category of a missing word, but might predict the wrong instance in that category (e.g., depending on the phrasing, BERT sometimes predicts *red* as the color of a dove). Finally, (iii) *limited reasoning capabilities*: it is unclear that LMs are capable of performing multiple reasoning steps involving implicit knowledge.

To increase the coverage of high-precision world knowledge and facilitate multi-hop reasoning by making intermediate reasoning steps explicit, prior work incorporated KBs (e.g. ConceptNet; Speer and Havasi, 2012) and knowledge-informed models into LM-based models (Xia et al., 2019; Bosselut and Choi, 2019; Chen et al., 2019).

In this paper, we study pre-trained LMs as an alternative to external KBs in providing knowledge

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 4615–4629,
November 16–20, 2020. ©2020 Association for Computational Linguistics



Because Brett found an internship while in college but Ian was unable to, **Brett** found a job less quickly after graduation. The purpose of the internship is to help people find jobs.

Because Brett found an internship while in college but Ian was unable to, **Ian** found a job less quickly after graduation. The purpose of the internship is to help people find jobs.

Because Brett found an internship while in college but Ian was unable to, **Brett** found a job less quickly after graduation. The definition of "job" is to be employed by someone.

Because Brett found an internship while in college but Ian was unable to, **Ian** found a job less quickly after graduation. The definition of "job" is to be employed by someone.

Figure 1: Model illustration for WinoGrande. Each answer choice (Brett, Ian) is assigned to the concatenation of the context and a clarification. The score for each choice is the best LM score across clarifications (2 in this case).

to commonsense question answering tasks. We propose an unsupervised model that uses an LM as the answer scorer, and a (possibly different) LM as a knowledge source. We formulate the process of obtaining relevant knowledge as a *self-talk*, inquiry-based discovery learning (Bruner, 1961), with the following steps: 1) seeking out knowledge by generating natural-language "clarification questions" conditioned on a given context, 2) generating their corresponding answers ("clarifications"), and 3) incorporating the clarifications as additional context.

Our model does not rely on external knowledge or additional supervision. Yet, we show that on 4 out of 6 tasks it substantially improves upon a zero-shot baseline that relies on LM score alone and performs on par, and sometimes better than, models that use external knowledge sources.

Integrating external knowledge warrants discerning relevant and helpful facts for solving a particular instance. LMs further require identifying that a clarification is factually-correct. We show that even among the clarifications that helped the prediction, humans perceived many as unhelpful or even incorrect, demonstrating that LM-based models often solve problems correctly for seemingly incorrect reasons. Our results call for future research on robust and correct knowledge integration to LM-based question answering systems.

2 Tasks

We focused on the multiple-choice question answering tasks detailed below. Each instance consists of an optional context, an optional question, and several answer choices.

COPA: Choice of Plausible Alternatives (Gordon et al., 2012): Asking about either a plausible cause or a plausible result, among two alternatives, of a certain event expressed in a simple sentence.

CommonSenseQA: commonsense Question Answering (Talmor et al., 2019): General questions about concepts from ConceptNet. To increase the challenge, the distractors are related to the target concept either by a relationship in ConceptNet or as suggested by crowdsourcing workers.

MC-TACO: Multiple Choice Temporal commonsense (Zhou et al., 2019): Questions about temporal aspects of events such as ordering, duration, frequency, and typical time. The distractors were selected in an adversarial way using BERT.[1]

Social IQa: Social Interaction Question Answering (Sap et al., 2019b): Questions regarding social interactions, based on the ATOMIC dataset (Sap et al., 2019a). Contexts describe social interactions and questions refer to one of a few aspects (e.g. the subject's motivation, following actions, etc.). The answers were crowdsourced.

PIQA: Physical Interaction Question Answering (Bisk et al., 2020): Questions regarding physical commonsense knowledge. Contexts are goals derived from an instruction website, typically involving less prototypical uses of everyday objects (e.g., using a bottle to separate eggs). The answers were crowdsourced, and an adversarial filtering algorithm was used to remove annotation artifacts.[2]

WinoGrande (Sakaguchi et al., 2020): A large-scale version of WSC that exhibits less bias thanks to adversarial filtering and use of placeholders instead of pronouns. As opposed to WSC that was curated by experts, WinoGrande was crowdsourced with a carefully designed approach that produces diverse examples which are trivial for humans.

3 Models

A given instance consists of an optional context c, an optional question q, and answer choices: $a_{i=1}^{k}$. We first describe the baseline model, which makes

[1] To make this task compatible with the other tasks, we only kept a single correct answer per instance, making our results not comparable to previously reported results.

[2] Word associations and dataset-specific features that are not informative for the task are identified by a strong baseline and removed (Gururangan et al., 2018; Zellers et al., 2018).

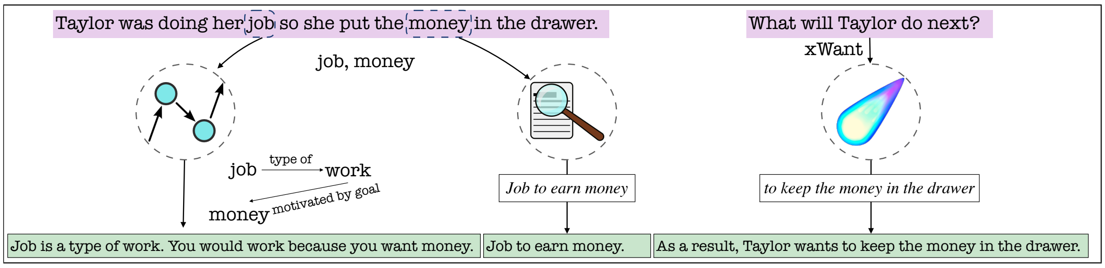

Taylor was doing her job so she put the money in the drawer.
What will Taylor do next?
xWant
job, money
job
type of
work
money
motivated by goal
Job to earn money
to keep the money in the drawer
Job is a type of work. You would work because you want money.
Job to earn money.
As a result, Taylor wants to keep the money in the drawer.

Figure 2: Generating a single clarification using ConceptNet, Google Ngrams, and COMeT (Social IQa instance).

the prediction based on the instance alone (§3.1). We then describe a knowledge-informed model that relies on external resources (§3.2). Finally, we discuss our self-talk model, which uses a pre-trained LMs to produce clarifications (§3.3).

3.1 LM-only Baseline

We use a pre-trained language model $\mathbb{LM}_s$ to score the plausibility of different text fragments. We experiment with the various LMs provided by the transformers package (Wolf et al., 2019): GPT (Radford et al., 2018), GPT2 (Radford et al., 2019, all sizes), a distilled GPT2 (Sanh et al., 2019), and XLNet (Yang et al., 2019, both sizes).

We assign each of the answer choices a_i into the combination of the context and the question, and obtain $\text{opt}_i = \text{combine}(c, q, a_i)$. The combine function is computed differently for each task. For example, in COPA, where the question might be either about the cause or the effect of the context, we create the following texts for cause: "[context]. *As a result,* [choice]" and for effect: "[context]. *The cause for it was that* [choice]".

We denote the score of each answer choice as $\text{score}(a_i) = \text{CE}(\text{opt}_i)$, where CE is cross-entropy loss defined as:
$\text{CE}(t_1...t_n) = -\frac{1}{n}\sum_{i=1}^{n} \log_2 p_{\mathbb{LM}_s}(t_i \mid t_1...t_{i-1})$. We predict the a_i with the lowest score as the correct answer, which is the most likely option according to $\mathbb{LM}_s$: $y = \text{argmin}_i \, \text{score}(a_i)$.

3.2 Baseline Model with External Knowledge

In the setup illustrated in Figure 1, each instance consists of an additional clarification list: $CL = \{cl_1, ..., cl_m\}$. Those are text fragments containing potentially relevant knowledge for solving the instance. For example, the clarification "*The purpose of the internship is to help people find jobs*" might help answering the question "*which of Brett and Ian found a job less quickly after graduation?*". We don't expect all the clarifications to be relevant and helpful for answering the main question. Instead, the model relies on the single clarification that increases its belief of a certain answer choice. Thus, the score of each answer choice is selected as the score of the text containing the clarification that most supports it, i.e., whose combination with it yields the minimal loss: $\text{score}(a_i) = \min_{cl \in CL} \text{CE}(opt_i + cl)$.
Again we predict $y = \text{argmin}_i \, \text{score}(a_i)$.

We extract clarifications from the following sources, exemplified in Figure 2.

ConceptNet. Similarly to previous work, we extract relation paths between words from the context and the question, and words from the answer choices. Since we incorporate the knowledge into the model as text, we convert each ConceptNet relation to a natural language template as in Davison et al. (2019). We limit the path length to 2 edges in order to maintain high precision.

Corpus. For pairs of words from the context and question and from the answer choices, we extract their joint occurrences (with minimum frequency of 100) in Google N-grams (Brants and Franz, 2006). This yields text fragments of up to 5 words rather than well-formed sentences, with the potential of describing the relationship between the two words (Shwartz and Dagan, 2018).

COMeT. COMeT (Bosselut et al., 2019) is a knowledge base construction model trained on the ATOMIC resource (Sap et al., 2019a) which consists of everyday situations along with multiple commonsense dimensions such as their causes, effects, pre- and post-conditions, etc. We generate all the dimensions unless we can generate specific relations that are more likely to help. Specifically, in Social IQa, we heuristically try to understand which type of relation in COMeT the question asks for. In COPA, we use the pre-condition relations for cause questions (`xIntent`, `xNeed`) and the post-condition relations for effect questions (`xEffect`,

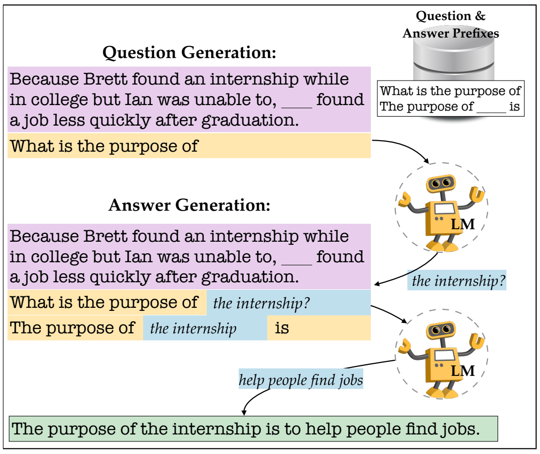

Question & Answer Prefixes
What is the purpose of
The purpose of ____ is
Question Generation:
Because Brett found an internship while in college but Ian was unable to, ____ found a job less quickly after graduation.
What is the purpose of
LM
Answer Generation:
Because Brett found an internship while in college but Ian was unable to, ____ found a job less quickly after graduation.
What is the purpose of the internship?
The purpose of the internship is
the internship?
LM
help people find jobs
The purpose of the internship is to help people find jobs.

Figure 3: Generating a clarification with LM: 1) Generate a question, conditioned on the context (pink) and question prefix (yellow). 2) Generate an answer, conditioned on the context, generated question and a corresponding answer prefix. The clarification is a concatenation of the answer prefix and generated text (green).

`xReact`, `xWant`, `oEffect`, `oReact`, `oWant`). When possible, we replace `personX` with the syntactic subject of the context or the question.

3.3 Self-talk Model

Our proposed model makes the prediction identically to Figure 1, but extracts the clarifications from pre-trained LMs. We treat the knowledge extraction from LMs as a process of self-asking clarification questions about the context and "discovering" their answers. Figure 3 exemplifies this process for WinoGrande with a generator language model $\mathbb{LM}_g$. For the sake of simplicity, the illustration depicts the process of generating a single pair of clarification question and answer.

We start by generating multiple clarification questions conditioned on the context, by 1) concatenating one of several question prefixes, which we curated for each task (e.g. "What is the purpose of", see Table 6 in the appendix); and 2) generating 5 questions for each prefix using Nucleus sampling with $p = 0.2$, i.e., sampling from the top 20% tokens (Holtzman et al., 2019).[3] We limit the question length to up to 6 additional tokens.

For each well-formed question that we obtained at the previous step, e.g. "*What is the purpose of the internship*?", we generate multiple answers using a similar method. Each question prefix corresponds to an answer prefix. We use the concatenation of the context, generated clarification question, and answer prefix as the prompt for generating an answer (clarification). We limit the answer length to 10 generated tokens, and use Nucleus sampling with $p = 0.5$. We generate 10 answers for each clarification question and keep all the well-formed clarifications. Note that the clarification questions themselves are only means to generate the clarifications, and they are not used by our model.[4]

Since we did not train the clarification generator to ask sensical, relevant, and helpful questions, nor did we train the answer generator to generate coherent and factually correct answers, we can assume that some of the generated clarifications do not provide useful information to the model.

4 Results

Table 2 displays the performance of the best model in each category according to the development accuracy. We report the performance of the following models: majority baseline, LM baseline (Baseline), LM-based model with external knowledge (Ext. Knowledge), Self-talk, supervised models from prior work when applicable (Pre. Sup), and human performance. Our zero-shot models are highlighted in purple. As expected, the overall performance is worse for the zero-shot models compared to the state-of-the-art supervised models, but they perform substantially better than the majority baselines on most tasks, with the exception of WinoGrande where they only slightly outperform it. Among the LM-based models, self-talk performs on par or within a few points from the external knowledge model.

Best Knowledge Source. Among the knowledge informed models, COMeT achieves the best performance across tasks. This likely happens because COMeT can dynamically generate predictions for any context, while the other two knowledge sources are static and lack coverage.

Table 1 shows the relative improvement in accuracy points compared to the zero-shot baseline,

[3] $p = 0.2$ is significantly lower than the standard value of $p = 0.9$ in the literature. We optimized for factual correctness, and our preliminary experiments have shown that lower p values produce texts that are more faithful to the LM training corpus, at the price of being more bland.

[4] In some datasets, an instance consists of a question. In this case, we can use the instance question as a "clarification" question and generate additional clarification questions similar to it. For example, the Social IQa context "*Austin fought for Quinn's life, but they eventually died on the operating table.*", the LM answers the question "*Why did Austin do this?*" directly with: "*Austin did this because they wanted to keep him alive*" (the correct answer is "*Because Austin wanted to save Quinn*").

	COMeT	ConceptNet	Google Ngrams	GPT	Distil-GPT2	GPT2	GPT2-M	GPT2-L	GPT2-XL	XLNet	XLNet-L
COPA	10.25	6.87	7.50	7.25	5.37	7.12	7.37	4.37	7.75	6.87	7.37
CSQA	0.39	-3.23	-0.30	-4.04	-3.79	-3.58	-3.09	-3.26	-3.65	-3.91	-3.55
MC-TACO	1.90	3.35	3.53	2.36	2.59	3.15	2.56	3.06	2.92	1.84	1.75
Social IQa	2.74	1.21	1.49	1.71	1.87	1.66	1.75	1.95	2.24	1.74	1.79
PIQA	3.77	4.07	4.36	4.01	3.61	3.80	3.89	3.88	3.96	3.82	4.10
WinoGrande	0.01	-0.01	-0.11	0.13	-0.17	-0.03	-0.04	0.04	0.08	-0.10	-0.25

Table 1: Relative improvement upon the zero-shot baseline in terms of development accuracy, for each knowledge source averaged across LMs for each dataset.

Dataset	**Model**	**LM**	**Knowledge Source**	**Dev Acc.**	**Test Acc.**
COPA	Majority			55.0	
	Baseline	Distil-GPT2		53.0	
	Ext. Knowledge	GPT2-L	COMeT	69.0	
	Self-talk	Distil-GPT2	Distil-GPT2	66.0	
	Pre. Sup	T5			94.8
	Human				100.0
Common SenseQA	Majority			20.9	
	Baseline	GPT-L		37.2	34.0
	Ext. Knowledge	GPT-XL	COMeT	39.7	36.2
	Self-talk	GPT-L	GPT-M	32.4	26.9
	Pre. Sup	Albert ensemble		83.7	76.5
	Human			88.9	88.9
MC TACO	Majority			40.3	43.0
	Baseline	GPT2-M		53.1	50.6
	External Knowledge	GPT2-XL	COMeT	58.8	55.6
	Self-talk	GPT2-XL	GPT2-XL	59.9	58.0
Social IQa	Majority			33.6	33.7
	Baseline	GPT2-L		41.1	41.1
	COMeT-CGA*		COMeT	49.6	51.9
	Ext. Knowledge	GPT2-XL	COMeT	47.5	45.3
	Self-talk	GPT2-XL	GPT2-L	46.2	43.9
	Pre. Sup	RoBERTa-large		76.6	77.1
	Human			86.9	84.4
PIQA	Majority			50.5	50.4
	Baseline	GPT2-XL		62.6	63.4
	Ext. Knowledge	GPT2-XL	COMeT	69.6	68.4
	Self-talk	GPT2-XL	GPT2-M	70.2	69.5
	Pre. Sup	RoBERTa-large		79.2	77.1
	Human			94.9	94.9
Wino Grande	Majority			50.4	50.4
	Baseline	GPT2-XL		54.8	54.8
	Ext. Knowledge	GPT2-XL	COMeT	55.4	53.7
	Self-talk	GPT2-XL	GPT	54.7	55.1
	Pre. Sup**	T5		86.5	84.6
	Human			94.1	94.0

Table 2: Best setup for each model type, according to development accuracy (excluding unpublished leaderboard submissions). Test accuracy is reported when labels are available or leaderboard submission was possible. *COMeT-CGA (Bosselut and Choi, 2019) is a zero-shot model performing probabilistic inference over generated inferences from a COMeT model trained on GPT2. ** (Lin et al., 2020).

for each knowledge source averaged across LMs for each dataset. Interestingly, the relative improvement is fairly uniform across knowledge sources, but it varies substantially across tasks. While some tasks benefit from any added knowledge, others benefit from none.

We also experimented with combining the clarifications from all the knowledge sources, which didn't prove beneficial except for MC-TACO (where it added +7.9 points to the dev accuracy, bringing it to 66.7). We assume that some resources added noise, making the whole smaller than the sum of its parts.

5 Analysis

While the performance on the end task serves as an extrinsic evaluation for the quality of the generated clarifications, we are also interested in evaluating it intrinsically. From preliminary experiments we know that there is a high ratio of noisy clarifications. We thus focus on and analyze two types of clarifications: useful (§5.1) and harmful (§5.2).[5]

5.1 Useful Clarifications

We define a clarification as *useful* if (a) it is the clarification with the best LM score in its instance (i.e., the clarification used in practice); and (b) the instance was incorrectly predicted by the zero-shot baseline but correctly predicted by the self-talk model. We sampled up to 50 useful clarifications for each combination of task and knowledge source, using the best performing LM (See Table 3 in the appendix for examples). We showed crowdsourcing workers an instance along with a clarification question and its answer, and asked them: 1) whether the question is grammatical, not entirely grammatical but understandable, or completely not understandable; and if the answer was anything but "completely not understandable", 2) whether the question is relevant, i.e. on topic with the instance. We asked the same questions about the answer, in addition to: 3) whether the answer is factually correct or likely true; and 4) whether the answer adds helpful information to solve the instance.

The annotation task was carried out in Amazon Mechanical Turk. To ensure the quality of annotations, we required that the workers be located in the US, UK, or Canada, and have a 99% approval rate for at least 5,000 prior tasks. We aggregated annotation from 3 workers using majority vote. The annotations yielded moderate levels of agreement, with

[5] We omitted COPA from the analysis due to its small size.

	COMET	ConceptNet	Distil-GPT2	GPT2	GPT2-M	GPT2-XL	GPT2-L	GPT	XLNet	XLNet-L
WinoGrande	72.00	43.80	36.00	61.20	83.00	68.00	71.10	67.90	72.70	83.30
Social IQa	90.00	56.00	66.00	74.00	72.00	76.00	76.00	80.00	36.00	52.00
MC-TACO	66.00	12.50	26.30	46.80	62.00	56.00	54.00	43.80	50.00	33.30
PIQA	72.00	40.00	38.00	62.00	72.00	60.00	66.00	35.00	75.00	33.30
CSQA	66.00	55.20	44.40	48.70	66.00	72.00	64.00	100.00	-	48.10
WinoGrande	60.00	43.80	40.00	24.50	46.80	46.00	53.30	39.30	45.50	33.30
Social IQa	76.00	42.00	28.00	48.00	36.00	42.00	50.00	50.00	22.00	28.00
MC-TACO	60.00	12.50	42.10	46.80	48.00	60.00	54.00	29.20	40.60	33.30
PIQA	62.00	44.00	24.00	44.00	44.00	42.00	36.00	0.00	50.00	33.30
CSQA	48.00	86.20	50.00	51.30	54.00	62.00	58.00	80.00	-	51.90
WinoGrande	34.00	12.50	20.00	14.30	34.00	24.00	31.10	35.70	27.30	33.30
Social IQa	-	20.00	-	-	-	-	-	-	-	-
MC-TACO	20.00	0.00	15.80	23.40	30.00	42.00	32.00	31.20	18.80	33.30
PIQA	28.00	6.00	14.00	16.00	30.00	26.00	24.00	5.00	25.00	33.30
CSQA	30.00	34.50	33.30	25.60	46.00	50.00	42.00	80.00	-	37.00

Figure 4: Ratio of clarifications considered as **relevant** (top), **factually correct** (middle), and **helpful** (bottom), among the useful and grammatical or understandable clarifications for each task and knowledge source. Answers in Social IQa were evaluated for helpfulness when the clarification question was different from the main question.

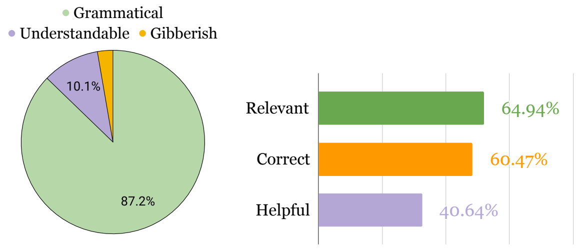


Figure 5: Human evaluation of the clarifications, aggregated across tasks and knowledge sources. **Left:** ratio of grammatical, not entirely grammatical but understandable, and completely not understandable clarifications. **Right:** percent of grammatical/understandable clarifications considered relevant, correct, and helpful.

Fleiss' Kappa $\kappa = 0.43$ (Landis and Koch, 1977). Among the different categories of annotations we measured pairwise accuracy, which ranged from 60.41% (the answer is factually correct) to 92.26% (the question is completely not understandable).

For the sake of brevity, we focus on the analysis of the answers to the clarification questions. The left part of Figure 5 shows that across tasks and resources, most clarifications are grammatical or at least understandable. Among the clarifications considered grammatical or understandable, the right part of the figure shows the percentage of clarifications considered relevant, correct, and helpful. Most clarifications were considered relevant to the context and factually correct, but only 40% on average were considered helpful. Considering that these are all clarifications that indeed helped the model, this is an interesting though not completely unexpected finding: the model utilizes knowledge that humans wouldn't consider as helpful.[6]

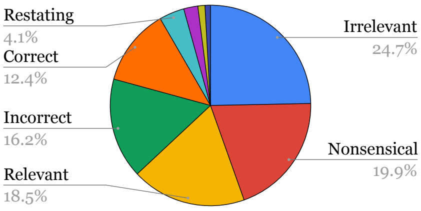


Figure 6: Types of errors caused by the harmful clarifications across all tasks and knowledge sources.

Breaking down by knowledge source, Figure 4 shows the ratio of clarifications considered by humans as **relevant** (top), **factually correct** (middle), and **helpful** (bottom), for each task and knowledge source. XLNet performs worse on all measures. ConceptNet's clarifications are often judged as irrelevant likely because they are limited to a very specific type of clarification (the relationship between a pair of terms). It's not too surprising that clarifications generated by LMs were sometimes judged as factually incorrect. We also note that COMeT generated factually correct clarifications for Social IQa (which is based on ATOMIC, on which COMeT was trained), and ConceptNet generated factually correct clarifications for CommonSenseQA (which is based on ConceptNet).

Table 3 demonstrates the types of knowledge in useful and relevant clarifications, showing that pre-trained LMs do particularly well in definitions.

[6] Seemingly unhelpful clarifications may yet increase the LM score by adding relevant lexical cues. A manual examination of a sample of answers judged as relevant but unhelpful revealed that 53.33% were answers for unhelpful questions, 20% were correct but unhelpful, 16.67% were factually incorrect, 10% were helpful to some extent (containing knowledge deemed too trivial by the annotators), and 10% had corresponding unanswerable instances.

Task	Source	Instance	Clarification
		Preconditions	
CSQA	ConceptNet	**Q:** Working on the elaborate task was taxing, it require extreme what? **Choices**: holding, concentration, energy, job, energy	**Q:** What is the relationship between 'working' and 'concentration'? **A:** In order for working to happen, concentration needs to happen.
	COMeT	**Q:** When you travel you should what in case of unexpected costs? **Choices**: go somewhere, energy, spend frivolously, fly in airplane, have money	**Q:** What do they want as a result? **A:** As a result, they want to make sure they have enough money.
MC-TACO	GPT2-XL	**C:** [...] Mark Bailey didn't dream of becoming a judge. **Q:** How many years did it take for Mark to become a judge? **Choices**: 63 years, [...], 7 months, 180, 7 weeks, 7 years, [...], 7 hours	**Q:** What happened first? **A:** The first thing that happened was he went to law school and became an attorney.
Social IQa	GPT2-M	**C:** Carson was at a friends house and decided to play video games. **Q:** What does Carson need to do before this? **Choices:** know about their friends house, of though about playing video games, of thought about their friend	**Q:** What does Carson need to do before this? **A:** Before doing that, Carson first had to get his mother's permission before playing the game.
		Social Commonsense	
Wino Grande	COMeT	**C:** Neil looked all over the neighborhood for their missing dog until Samuel called and __ got their dog back. **Choices**: Neil, Samuel	**Q:** What does Neil feel as a result? **A:** As a result, Neil feels happy to have the dog back.
		Definitions and Lexical Knowledge	
PIQA	COMeT	**Q:** how to sanitize a toothbrush **Choices**: [...] antibacterial mouthwash into a small toilet bowl [...] soak your toothbrush [...] [...] antibacterial mouthwash into a small cup [...] soak your toothbrush [...]	**Q:** What does you want as a result? **A:** As a result, you wants to get rid of germs.
Wino Grande	GPT2-L	**C:** During the summer, I like visiting the zoo more than the aquarium because the __ is inside. **Choices**: zoo, aquarium	**Q:** What is the definition of 'zoo'? **A:** The definition of a 'zoo' is an area where people can see animals in captivity.
Social IQa	ConceptNet	**C:** Jan had been on vacation in Hawaii for a week. Last night, she finally returned home. **Q:** What will Jan want to do next? **Choices**: return to her family, Return to work, Leave for another vacation.	**Q:** What is the relationship between 'vacation' and 'work'? **A:** Vacation is the opposite of work.
		Functions	
CSQA	DistilGPT2	**Q:** What do professors primarily do? **Choices**: master physics, state facts, wear wrinkled tweed jackets, school students, teach courses	**Q:** What is the main function of a professor's teaching career,? **A:** The main function of a professor's teaching career, is to teach students how they can improve their knowledge.

Table 3: Example of relevant useful clarifications and the type of knowledge they provide.

5.2 Harmful Clarifications

Symmetrically, we also study the *harmful* clarifications. A clarification is harmful if (a) it is the clarification with the best LM score in its instance; and (b) the instance was correctly predicted by the zero-shot baseline but incorrectly predicted by the self-talk model. We sampled up to 25 harmful clarifications from the predictions of the best setup (LM and knowledge source) for each task, and manually categorized the errors into the following types.

1. **Irrelevant**: the clarification was off topic.
2. **Nonsensical or ungrammatical**: the clarification was not a complete sentence, or had other grammar or meaning issues.
3. **Relevant**: the clarification contributed relevant knowledge but it wasn't enough for predicting the correct answer.
4. **Factually Incorrect**: the clarification made a factually incorrect statement, often in support of one of the distractors.
5. **Correct**: the clarification yielded an alternative correct answer for the main instance.
6. **Restating the instance**: the clarification repeated the context or the main question.
7. **Wrong sense**: the clarification interpreted a word from the instance in the wrong sense.
8. **Dataset error**: the instance is incorrect or lacks information required for answering it correctly.

Figure 6 shows the percent of each error type across all the tasks and knowledge sources. The majority of clarifications are irrelevant, ungrammatical or nonsensical, or relevant but not helpful for making the correct prediction. We judged a non-negligible 12.4% of the clarifications as providing alternative correct answers, phrased differently from the gold answer. Table 4 provides an instance for each error type.[7]

6 Related Work

6.1 External Knowledge in Neural Models

Approaches for incorporating external knowledge into a neural model consist of several components: (1) the task addressed; (2) neural model; (3) knowledge sources; and (4) incorporation method. Most models target tasks that require commonsense knowledge, such as the story cloze test (Mostafazadeh et al., 2016) and machine comprehension tasks (Kočiskỳ et al., 2018; Ostermann

[7] See Figures 8 and 9 in the appendix for a breakdown of error types by task and knowledge source.

Task	Know. Source	Instance	Clarification
Irrelevant			
PIQA	GPT2-XL	**Q:** how do you sit a baby in a restaurant? **Choices**: place them in a booster seat., place them on the table.	**Q:** What is the definition of "a good time"? **A:** The definition of "a good time" is not the same as what constitutes an acceptable meal.
Nonsensical or ungrammatical			
Social IQa	XLNet	**C:** Cameron went out of their way to help a friend who was in need of help. **Q:** How would Cameron feel after helping? **Choices**: Cameron then frustrated., Cameron then happy., Cameron then annoyed.	**Q:** How would Cameron feel after helping? **A:** Cameron felt they were doing, but then he realized that.
Relevant			
WinoGrande	GPT2-XL	The children were not vaccinated, which was fine with Betty but annoyed Mary. ___ believed they made kids autistic. **Choices**: Betty, Mary	**Q:** What does it mean to be "autistic"?" **A:** Be "autistic" means to have problems in social interaction and communication skills.
Factually Incorrect			
WinoGrande	COMeT	Derrick really liked working in Human Resources and William wanted to work at the same company, and ___ subsequently offered a position. **Choices**: Derrick, William	**Q:** What will Derrick want as a result? **A:** As a result, Derrick wants to go to work with them
Correct			
CSQA	Google Ngrams	**Q:** What do people usually feel when falling in love? **Choices**: getting married, pain, happiness, getting married, suffering	**Q:** - **A:** Suffering from unrequited love.
Restating the instance			
CSQA	COMeT	**Q:** Billy set aside a block of time for having fun after work. Why might he do this? **Choices**: happiness, stress relief, pleasure, ocean, may laugh	**Q:** What will Billy want as a result? **A:** As a result, they want to do something fun.
Wrong Sense			
MC-TACO	ConceptNet	**C:** [...] Islam thrived as a strong, male-dominated religion of individuality [...] preaching brotherhood [...]. **Q:** What happened after Islam became popular in the region? **Choices**: they drank liquor, it died off, it expanded even further, they drank alcohol, it died out, it died down	**Q:** What is the relationship between brotherhood and alcohol? **A:** You are likely to find brotherhood in a fraternity house. You are likely to find alcohol in a fraternity house.

Table 4: An example for each of the error types among the harmful clarifications.

et al., 2018; Clark et al., 2018; Talmor et al., 2019). The neural component has recently shifted from biLSTM to transformer-based representations, specifically pre-trained LMs (Devlin et al., 2019; Liu et al., 2019).

With respect to the knowledge source, the vast majority of papers rely on ConceptNet to extract relation paths between concepts and entities identified in the input (Speer and Havasi, 2012, see an example in Figure 2). Additional resources include WordNet (Lin et al., 2017; Wang and Jiang, 2019), retrieval or statistics mind from corpora (Lin et al., 2017; Mitra et al., 2019; Joshi et al., 2020), knowledge base embeddings (Chen et al., 2019; Xiong et al., 2019), hand-crafted rules (Lin et al., 2017; Tandon et al., 2018), and tools such as sentiment analyzers (Chen et al., 2019) and knowledge-informed LMs (Bosselut and Choi, 2019).

The external knowledge is typically incorporated into the neural model by learning a vector representation of the symbolic knowledge (e.g. subgraphs from ConceptNet), and attending to it via attention mechanism when representing the inputs (Bauer et al., 2018; Paul and Frank, 2019; Lin et al., 2019). Alternative approaches include using the knowledge to score answer candidates and prune implausible ones (Lin et al., 2017; Tandon et al., 2018), and training in a multi-task setup via auxiliary tasks pertaining to knowledge (Xia et al., 2019).

To the best of our knowledge, our method is the first to generate knowledge from pre-trained language models and incorporate it as external knowledge into a question answering model. Concurrently, Latcinnik and Berant (2020) used one language model to generate hypotheses and another language model as an answer scorer for CommonSenseQA.

6.2 Extracting Knowledge from LMs

Pre-trained LMs such as GPT2 (Radford et al., 2019) and BERT (Devlin et al., 2019) capture various types of world knowledge. Petroni et al. (2019) showed that such LMs can be used in a KB completion task over ConceptNet and Wikidata (Vrandečić and Krötzsch, 2014) by converting KB relations into natural language templates and querying the LM for the missing part in the triplet (concept$_1$, relation, concept$_2$). For instance, querying BERT for suitable substitutes to the mask in "Dante was born in [MASK]" assigns the highest probability to Rome. Davison et al. (2019) similarly showed that BERT assigns higher scores to natural language fragments of true rather than fictitious ConceptNet triplets, and semi-automated the template creation by using GPT2 to score hand-crafted templates.

While both works have shown somewhat promising results, other work showed that knowledge extracted from LMs is expectantly not always accurate. Specifically, Kassner and Schütze (2020) showed that negated facts are also considered likely by the LM, while Logan et al. (2019) pointed out that LMs may over-generalize and produce incorrect facts such as "Barack Obama's wife is Hillary".

6.3 Generating Questions and Explanations

There are numerous research directions investigating automatic question generation (Vanderwende, 2008). Motivations vary from data augmentation to QA tasks (Du et al., 2017; Dhingra et al., 2018; Du and Cardie, 2018; Sachan and Xing, 2018; Fabbri et al., 2020) through conversational machine reading (Saeidi et al., 2018; Pan et al., 2019), simplifying questions to make them more easily answerable (Buck et al., 2018; Talmor and Berant, 2018; Perez et al., 2020), to using questions as means for other purposes such as sentence representation and summarization (Guo et al., 2018; Potash and Suleman, 2019).

In particular, our work is pertinent to previous work in producing clarification questions and explanations. Rao and Daumé III (2019) worked on questions from forums (e.g. Stack Exchange). They proposed a model that generates clarification questions and corresponding answers for a given question, using the question's comments (clarification questions and answers) as supervision. Question-answer pairs were scored based on how much relevant information they add to the context.

Shen et al. (2019) developed an active learning framework for image captioning that learns to detect uncertainty about generated words and ask natural language questions to reduce its uncertainty. A visual question answering (VQA) model provides an answer which is then used to change the caption. The framework is trained with reinforcement learning, but the gold standard captions are used during a warmup steps and the VQA model is supervised.

Klein and Nabi (2019) proposed a joint question generation and question answering framework. They fine-tuned GPT2 on a question answering dataset to generate a question and an answer span for a given passage, and trained BERT to answer the generated question given the passage. Finally, Rajani et al. (2019) proposed a model for CommonSenseQA that generates explanations for its predictions. They collected human explanations and used them to fine-tune LMs to automatically generate explanations. These explanations were then added as additional inputs. The shortcoming of this approach is that it requires collecting specific human explanations for each new dataset.

7 Discussion and Conclusion

We presented an unsupervised framework for multiple choice commonsense tasks that generates and integrates background knowledge from pre-trained LMs. On most tasks, it performs substantially better than the baseline and similarly to a model that had access to external knowledge resources.

We have listed several shortcomings of using pre-trained LMs as knowledge providers: (i) *insufficient coverage*, (ii) *insufficient precision*, and (iii) *limited reasoning capabilities*. Despite their insufficient precision compared to a KB like ConceptNet, we showed that clarifications generated by LMs resulted in similar or superior empirical gains. Among the clarifications used in practice by the answer scorer, about 60% of those that yielded a correct prediction and 12% of those that yielded an incorrect prediction were judged by humans as factually correct.

By design, our model makes a single additional reasoning step explicit, aiming to facilitate reasoning about implicit inferences. A preliminary experiment in which we incorporated clarification pairs to facilitate two hops got mixed results. An interesting future direction is to generate each clarification in response to the previous ones, in a dialogue setup (Saeidi et al., 2018). Another challenge is the "needle in a haystack" problem of the clarifications, and one way to address it is to develop a model that is capable of "introspection", specifically knowing what it doesn't know. A more structured knowledge generation might also make the combination of various knowledge sources more successful.

Filling in knowledge gaps and making implicit intermediate reasoning steps explicit is imperative going forward. We hope that our framework will facilitate future research in this area. Our code and data will be made available upon publication. Our code and data is available at github.com/vered1986/self_talk.

Acknowledgements

This research was supported in part by NSF (IIS-1524371, IIS-1714566), DARPA under the CwC program through the ARO (W911NF-15-1-0543), and DARPA under the MCS program through NIWC Pacific (N66001-19-2-4031).

References

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230, Brussels, Belgium. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Antoine Bosselut and Yejin Choi. 2019. Dynamic knowledge graph construction for zero-shot commonsense question answering. *ArXiv*, abs/1911.03876.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1 (2006). *Linguistic Data Consortium, Philadelphia*.

Jerome S Bruner. 1961. The act of discovery. *Harvard educational review*, 31:21–32.

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. 2018. Ask the right questions: Active question reformulation with reinforcement learning. In *International Conference on Learning Representations*.

Jiaao Chen, Jianshu Chen, and Zhou Yu. 2019. Incorporating structured commonsense knowledge in story completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6244–6251.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Joe Davison, Joshua Feldman, and Alexander 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, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. 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 582–587, New Orleans, Louisiana. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2018. Harvesting paragraph-level question-answer pairs from Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1907–1917, Melbourne, Australia. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352, Vancouver, Canada. Association for Computational Linguistics.

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513, Online. Association for Computational Linguistics.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 687–697, Melbourne, Australia. 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.

Lynette Hirschman, Marc Light, Eric Breck, and John D Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 325–332. Association for Computational Linguistics.

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

Mandar Joshi, Kenton Lee, Yi Luan, and Kristina Toutanova. 2020. Contextualized representations using textual encyclopedic knowledge. *arXiv preprint arXiv:2004.12006*.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Tassilo Klein and Moin Nabi. 2019. Learning to answer by learning to ask: Getting the best of gpt-2 and bert worlds. *arXiv preprint arXiv:1911.02365*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

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

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *arXiv preprint arXiv:2004.05569*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Hongyu Lin, Le Sun, and Xianpei Han. 2017. Reasoning with heterogeneous knowledge for commonsense machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2032–2043, Copenhagen, Denmark. Association for Computational Linguistics.

Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2020. Tttttackling winogrande schemas.

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*.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's Wife Hillary: Using Knowledge Graphs for Fact-Aware Language Modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Arindam Mitra, Pratyay Banerjee, Kuntal Kumar Pal, Swaroop Mishra, and Chitta Baral. 2019. How additional knowledge can improve natural language commonsense question answering?

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. Semeval-2018 task 11: Machine comprehension using commonsense knowledge. In *Proceedings of the 12th International Workshop on semantic evaluation*, pages 747–757.

Boyuan Pan, Hao Li, Ziyu Yao, Deng Cai, and Huan Sun. 2019. Reinforced dynamic reasoning for conversational question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2124, Florence, Italy. Association for Computational Linguistics.

Debjit Paul and Anette Frank. 2019. Ranking and selecting multi-hop knowledge paths to better predict human needs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3671–3681, Minneapolis, Minnesota. Association for Computational Linguistics.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. In *RCQA workshop @ AAAI 2020*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Peter Potash and Kaheer Suleman. 2019. Playing log (n)-questions over sentences. In *EmeCom workshop @ NeurIPS 2019*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. -.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2019. Answer-based Adversarial Training for Generating Clarification Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 143–155, Minneapolis, Minnesota. Association for Computational Linguistics.

Mrinmaya Sachan and Eric Xing. 2018. Self-training for jointly learning to ask and answer questions. 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 629–640.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WINOGRANDE: An adversarial winograd schema challenge at scale. In *AAAI*.

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

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

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Tingke Shen, Amlan Kar, and Sanja Fidler. 2019. Learning to caption images through a lifetime by asking questions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10393–10402.

Vered Shwartz and Ido Dagan. 2016. Path-based vs. distributional information in recognizing lexical semantic relations. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex - V)*, pages 24–29, Osaka, Japan. The COLING 2016 Organizing Committee.

Vered Shwartz and Ido Dagan. 2018. Paraphrase to explicate: Revealing implicit noun-compound relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1200–1211, Melbourne, Australia. Association for Computational Linguistics.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. 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 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

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

Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut, and Peter Clark. 2018. Reasoning about actions and state changes by injecting commonsense knowledge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 57–66, Brussels, Belgium. Association for Computational Linguistics.

Raphael Tang and Jimmy Lin. 2018. Adaptive pruning of neural language models for mobile devices. *arXiv preprint arXiv:1809.10282*.

Lucy Vanderwende. 2008. The Importance of Being Important: Question Generation. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Chao Wang and Hui Jiang. 2019. Explicit utilization of general knowledge in machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2263–2272, Florence, Italy. Association for Computational Linguistics.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? a pilot study for sense making and explanation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4020–4026, Florence, Italy. 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.

Jiangnan Xia, Chen Wu, and Ming Yan. 2019. Incorporating relation knowledge into commonsense reading comprehension with multi-task learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2393–2396.

Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Improving question answering over incomplete KBs with knowledge-aware reader. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4258–4264, Florence, Italy. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3354–3360.

A Question and Answer Prefixes

We came up with question and answer prefixes by experimenting with a few generic prefixes and observing what generally yields accurate answers. For example, we observed that LMs are not very good at causal and temporal relationships but are pretty good at definitions. For the datasets whose instances include questions (e.g. Social IQa) we also used the corresponding question prefixes.

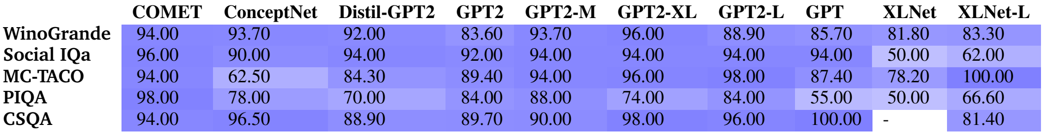
Table 6 presents the question and answer prefixes used for each task. "_" in the answer prefix is replaced with the generated question (excluding the question mark), e.g. "What is the definition of a cat?" yields the answer prefix: "The definition of a cat is". The Social IQa templates correspond to COMeT dimensions. X is replaced with the syntactic subject of the sentence.

B Best Language Model

Table 5 shows the average development accuracy of the LMs across the different knowledge sources. In general there is a preference to GPT-2, and in particular to the larger models, except for COPA in which the distilled version works best. A possible explanation might be that the language model distillation reduces the likelihood of rare words (Tang and Lin, 2018), which works well for the simple sentences in COPA. The XLNet models perform poorly, perhaps due to their smaller training corpus (16GB vs 40GB in GPT-2, both using web text).

	GPT	Distil-GPT2	GPT2	GPT2-M	GPT2-L	GPT2-XL	XLNet	XLNet-L
COPA	58.64	63.73	59.73	61.82	60.64	57.91	51.91	49.45
CSQA	27.57	25.45	25.64	27.74	31.75	31.22	21.47	20.79
MC-TACO	47.72	48.75	50.06	52.99	56.61	58.05	34.18	37.03
Social IQa	41.62	40.39	41.80	43.39	44.39	45.50	33.12	33.65
PIQA	57.91	59.63	61.95	65.57	67.89	69.59	49.24	48.80
WinoGrande	52.18	50.94	51.16	50.18	52.85	54.04	49.07	48.74

Table 5: Average self-talk accuracy for each LM answer scorer, averaged across knowledge sources.



	COMET	ConceptNet	Distil-GPT2	GPT2	GPT2-M	GPT2-XL	GPT2-L	GPT	XLNet	XLNet-L
WinoGrande	94.00	93.70	92.00	83.60	93.70	96.00	88.90	85.70	81.80	83.30
Social IQa	96.00	90.00	94.00	92.00	94.00	94.00	94.00	94.00	50.00	62.00
MC-TACO	94.00	62.50	84.30	89.40	94.00	96.00	98.00	87.40	78.20	100.00
PIQA	98.00	78.00	70.00	84.00	88.00	74.00	84.00	55.00	50.00	66.60
CSQA	94.00	96.50	88.90	89.70	90.00	98.00	96.00	100.00	-	81.40

Figure 7: Ratio of clarifications considered by humans as **grammatical or understandable** among the useful clarifications for each task and knowledge source.

Dataset	Question Prefix	Answer Prefix
COPA & CSQA	What is the definition of What is the main purpose of What is the main function of a What are the properties of a What is a What happened as a result of What might have caused	The definition of _ is The purpose of _ is to The main function of a _ is The properties of a _ are that _ is As a result of _, The cause of _ was
MC TACO	How long did this take? How often does this happen? How many times did this happen? What happened first? What happened last?	This lasted for Every This happened The first thing that happened was The last thing that happened was
Social IQa	What will X want to do next? What will X want to do after? How would X feel afterwards? How would X feel as a result? How would X feel after? How would you describe X? What kind of person is X? How would you describe X as a person? Why did X do that? Why did X do this? Why did X want to do this? What does X need to do beforehand? What does X need to do before? What does X need to do before this? What did X need to do before this? What will happen to X? What will happen to X next? What will X do next? What did X do?	X wanted X wanted X felt X felt X felt X is a X is a X is a X did this because they wanted X did this because they wanted X did this because they wanted Before doing that, X first had to Before doing that, X first had to Before doing that, X first had to Before doing that, X first had to X X X What X did was
PIQA	How to How do you How can one What can be used for What can one do in order to What should you use for What is the definition of What are the properties of a What is a	The way to do _ is The way you do _ is One can _ by _ can be used for In order to _, one can For _, you should you use The definition of _ is The properties of a _ are that _ is
Wino Grande	What is the definition of What is the main purpose of What is the main function of a What are the properties of a What is What does it mean to	The definition of _ is The purpose of _ is to The main function of a _ is The properties of a _ are that _ is _ means

Table 6: Question & answer prefixes used for each task.

C Analysis

C.1 Useful Clarifications

Figure 7 shows, for each task and knowledge source, the ratio of useful clarifications that were considered by humans as either grammatical or at least understandable. The majority of the helpful clarifications are considered as grammatical. The XLNet models are slightly worse in terms of grammaticality. For example, the clarification question "*What are the properties of a you sharpen a pencil, ?*" and the answer "*The properties of a you sharpen a pencil, are that it will not break or be dulled*" generated for the PIQA instance "sharpen a pencil" by XLNet-base. Despite its grammar errors, the answer was still useful for a LM to determine the correct answer.

C.2 Harmful Clarifications

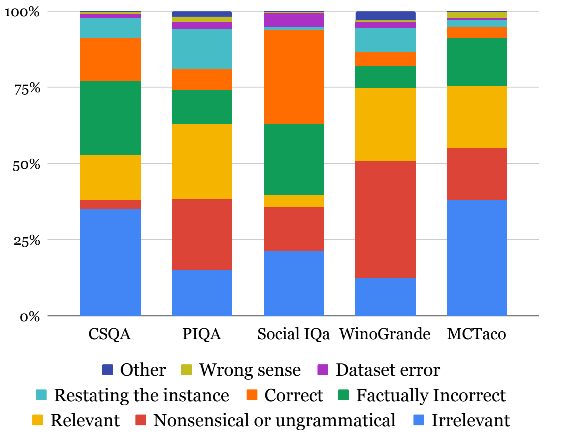


Figure 8: Types of errors caused by the harmful clarifications, for each task, across all knowledge sources.

Figure 8 breaks down by task the type of errors found in the harmful clarifications. In Social IQa and CommonSenseQA, many alternative correct answers are generated, but this doesn't happen in WinoGrande, that by design only allows for one correct answer. Clarifications in MC-TACO are more than average irrelevant. In the future, it would be interesting to investigate whether this is due to inherent lack of temporal commonsense in LMs or due to misguided attempts to extract it.

Figure 9 similarly breaks down the errors by

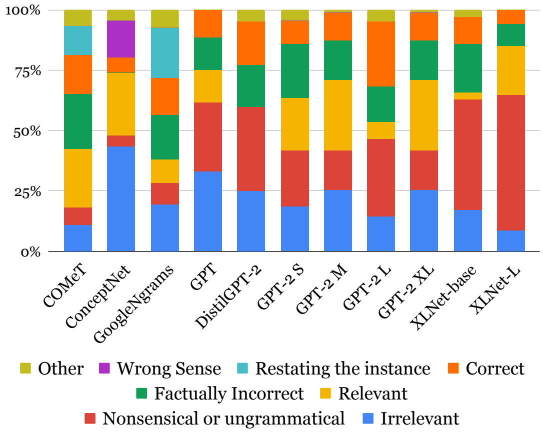


Figure 9: Types of errors caused by the harmful clarifications, for each knowledge source, across all tasks.

knowledge source. All knowledge sources except for ConceptNet make incorrect statements, but LMs also tend to make nonsensical statements, especially XLNet. ConceptNet tends to generate irrelevant clarifications (about the relationship between two unimportant terms). Being a static resource, is was also insensitive to the word senses. Google Ngrams, the only other static knowledge source, didn't suffer from this issue. This is likely because a polysemous term x related to y in one of its senses wouldn't typically co-occur with y in its non-related senses (Shwartz and Dagan, 2016).