Do Neural Language Models Overcome Reporting Bias?

Vered Shwartz and Yejin Choi
Allen Institute for AI
Paul G. Allen School of Computer Science & Engineering, University of Washington
{vereds,yejinc}@allenai.org

Abstract

Mining commonsense knowledge from corpora suffers from *reporting bias*, over-representing the rare at the expense of the trivial (Gordon and Van Durme, 2013). We study to what extent pre-trained language models overcome this issue. We find that while their generalization capacity allows them to better estimate the plausibility of frequent but unspoken of actions, outcomes, and properties, they also tend to overestimate that of the very rare, amplifying the bias that already exists in their training corpus.

1 Introduction

Apart from several notable efforts to collect commonsense knowledge from experts (Lenat, 1995) or through crowdsourcing (Speer and Havasi, 2012; Sap et al., 2019), most work has been on extracting such knowledge from large text corpora (Mitchell et al., 2018). While the latter approach is scalable and low cost, it also suffers from *reporting bias*: due to Grice's conversational maxim of quantity (Grice et al., 1975), people rarely state the obvious, thus many trivial facts ("people breathe") are rarely mentioned in text, while uncommon events ("people murder") are reported disproportionately (Gordon and Van Durme, 2013; Sorower et al., 2011).

Traditionally, knowledge acquisition from text was extractive. In recent years, the generalization capacity of neural language models (LMs) and their ability to aggregate knowledge across contexts have facilitated estimating the plausibility of facts, even when they don't appear in the corpus explicitly. Recent pre-trained LMs such as GPT-2 (Radford et al., 2019) and BERT (Devlin et al., 2019), trained on massive texts, dominate the NLP leaderboards, and are considered a source of commonsense knowledge (Petroni et al., 2019). Does this mean that pre-trained LMs overcome reporting bias?

In this paper we revisit the experiments conducted by Gordon and Van Durme (2013) (henceforth G&V), applying them to various pre-trained LMs (based on the nature of the experiment, we test either masked LMs or standard left-to-right LMs). We find that LMs, compared to extractive methods:[1]

1. Provide a worse estimate of action frequency, mostly due to overestimating very rare actions.
2. Predict both expected outcomes as well as sensational and unlikely outcomes.
3. Are capable of learning associations between concepts and their properties indirectly, but tend to over-generalization, which leads to confusing semantically-similar but mutually exclusive values.

2 Actions and Events

G&V demonstrate the discrepancy between corpus occurrences and actual action frequency by showing that if you believe the corpus, people murder more than they breathe. Breathing is an activity we take for granted and thus rarely talk about (Grice et al., 1975). That murder is frequent in the corpus is a reflection of the same issue: we talk more about uncommon or newsworthy events (van Dalen, 2012).

We follow G&V's qualitative analysis of actions and events performed by or which happen to people by comparing real-world frequency to corpus-based and LM-based frequency. We estimate real-world

[1] Our data and code are publicly available at https://github.com/vered1986/reporting_bias_lms.

Proceedings of the 28th International Conference on Computational Linguistics, pages 6863–6870
Barcelona, Spain (Online), December 8-13, 2020

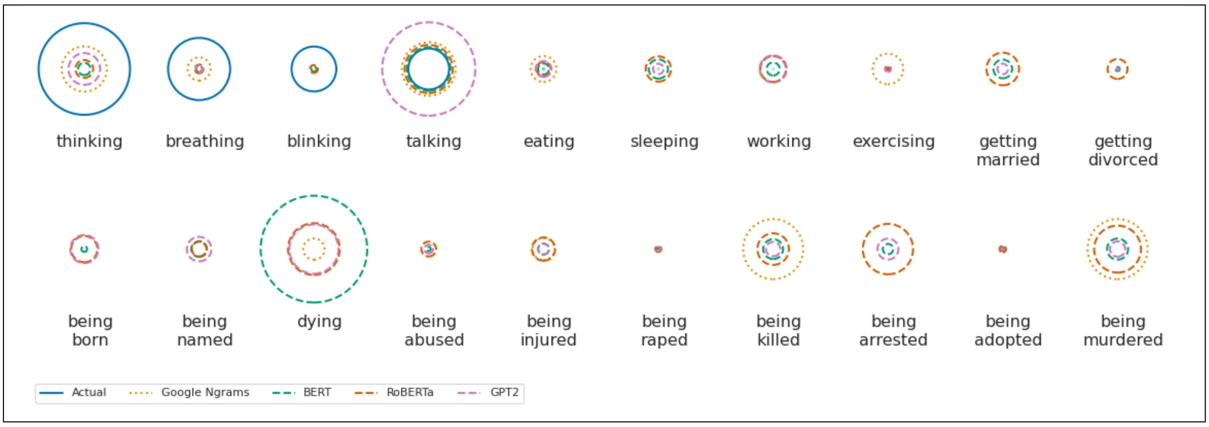

Figure 1: Frequency of actions performed or occurring to people during their lifetime from very frequent (daily), through once in a lifetime events, to very rare (don't happen to most people). Note that actual frequencies of rare events are too small to show. See Appendix A for the exact frequencies.

	BERT	**RoBERTa**	**GPT-2**		**BERT**	**RoBERTa**	**GPT-2**
The person ___.	wins (11.4) died (11.4) dies (10.6) won (7.8) lost (3.5) said (2.4) speaks (1.9) answered (1.6) replied (1.3) loses (1.3)	said (5.8) responds (4.0) replied (3.4) dies (3.3) died (2.9) responded (2.5) says (2.4) replies (2.2) asked (2.1) commented (2.1)	let (4.3) see (3.9) make (2.4) get (2.1) look (2.1) take (1.2) set (1.2) give (1.1) using (1.1) go (1.1)	The person is ___.	killed (7.5) married (6.6) dying (4.2) deceased (3.8) eliminated (2.6) retired (2.2) lost (2.0) arrested (2.0) elected (1.5) disabled (1.5)	gone (6.3) deceased (3.8) arrested (2.9) missing (2.5) responding (1.9) involved (1.9) reading (1.9) dying (1.9) confused (1.5) reporting (1.5)	let (4.3) see (3.9) make (2.4) get (2.1) look (2.1) take (1.2) set (1.2) give (1.1) using (1.1) go (1.1)

Table 1: Top LM predictions for actions performed by people along with their scores (percents).

frequency (e.g. *how many times does a person breathe in their lifetime?*) from published statistics based on US data, as detailed in Appendix A. Corpus frequency is computed using the Google N-gram corpus (Brants and Franz, 2006). Specifically, we compute the normalized frequency of the verbs appearing in the 3-gram "person is <verb>", falling back to the bigram "person <verb>" if no results are found. We use SpaCy to determine parts of speech, keeping non auxiliary verbs (Honnibal and Montani, 2017).

While LM scores don't represent frequency or probability, they are often used in practice as a proxy for plausibility. Thus, we would expect LM scores to correlate with real-world frequency. We query masked LMs for substitutes of the mask in several templates describing actions,[2] and left-to-right LMs by greedily decoding the next token (e.g. for "The person is"), taking the maximum score for each word across templates. Specifically, we use BERT large uncased (Devlin et al., 2019), RoBERTa large (Liu et al., 2019), and GPT-2 XL (Radford et al., 2019) from the Transformers package (Wolf et al., 2019). We keep the non auxiliary verbs among the top 5000 predictions.[3]

Figure 1 visualizes the relative frequency of each action as estimated by the various sources, where the scores for all actions are normalized for each source. Actions are sorted by their real-world frequency from very frequent to very rare. First, we observe that LMs assign non-zero scores for all actions, as opposed to the non-smoothed corpus frequencies from Google Ngrams. However, the scores they produce diverge further from the actual distribution, measuring with KL-divergence: Google Ngrams - 2.94, BERT and GPT-2 - 3.77, and RoBERTa - 3.08. LMs produce a more accurate estimate for some frequent actions (blinking, eating) but worse for others (thinking, breathing). At the same time, LMs also exaggerate the frequencies of rare events (e.g. dying), producing estimates not only higher than the actual frequency but even higher than the corpus frequency.

The same patterns emerge for both LMs, but some exceptions stand out. For example, BERT overestimates the frequency of dying, which may be due to being trained on Wikipedia, which consists of many entries describing historically important—and dead—people. RoBERTa, on the other hand, which

[2] "The person is [MASK].", "The person [MASK].", "People are [MASK].", "All people [MASK]."

[3] We consider some synonyms and subactions, e.g. including "exhale" and "inhale" in "breathe", as detailed in Appendix A.

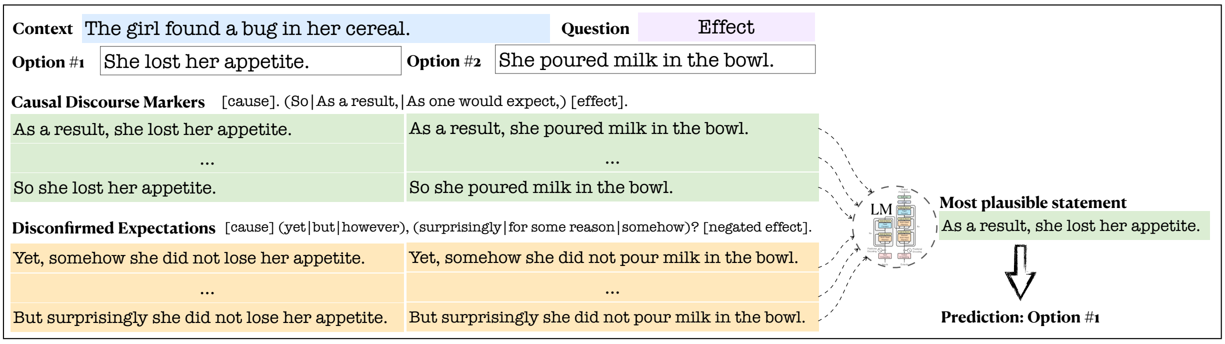
Context
The girl found a bug in her cereal.
Question
Effect
Option #1
She lost her appetite.
Option #2
She poured milk in the bowl.
Causal Discourse Markers
[cause]. (So|As a result,|As one would expect,) [effect].
As a result, she lost her appetite.
As a result, she poured milk in the bowl.
...
...
So she lost her appetite.
So she poured milk in the bowl.
Disconfirmed Expectations
[cause] (yet|but|however), (surprisingly|for some reason|somehow)? [negated effect].
Yet, somehow she did not lose her appetite.
Yet, somehow she did not pour milk in the bowl.
...
...
But surprisingly she did not lose her appetite.
But surprisingly she did not pour milk in the bowl.
LM
Most plausible statement
As a result, she lost her appetite.
Prediction: Option #1

Figure 2: Illustration of the Zero-shot+DE model prediction of an instance from COPA. Each answer choice has a set of support statements including causal discourse markers (both models) and disconfirmed expectations (only in Zero-shot+DE). The LM is used to score the statements for plausibility, and the model predicts the answer choice associated with the most plausible statement.

was trained on the web, overestimates the frequency of newsworthy events such as being murdered or arrested. Table 1 further exemplifies the top LM predictions for actions performed by people, using additional templates. While most predictions, especially by GPT-2, are common or mundane verbs (*said*), some describe rarer events (*killed*).

3 Event Outcomes

G&V argue that an event outcome is more likely to be mentioned in text if it's not certain. For instance, "The man turned on the faucet. The water started running in a steady stream." makes an awfully boring story, while "Water gushed out of the sink" builds up to a turn in events. Do LMs learn the proportional outcome distribution in the corpus, or can they overcome it by implicitly learned commonsense?

A good testbed for event outcomes is the COPA dataset (Choice of Plausible Alternatives) (Gordon et al., 2012). Given an event (context), the goal is to predict its cause or effect among two candidate answers. We focus on LMs typically used for generation: GPT (Radford et al., 2018), GPT-2 (Radford et al., 2019), and XLNet (Yang et al., 2019). Table 2 exemplifies outcomes predicted for several COPA events with various LMs and decoding strategies: top $k = 10$ (Fan et al., 2018), top $p = 0.9$ (Holtzman et al., 2020), and beam search with beam size of 5. We observe a combination of mundane, correct outcomes (water running in a steady stream) and sensational and unlikely events ("the fire broke out").

In order to quantify the ability of LMs to predict outcomes, we target the multiple choice COPA task with a zero-shot LM-based model (Zero-shot in Table 3). For a given context and for each candidate answer, we create a set of *supporting statements*: `[cause] [causal discourse marker] [effect]`, as exemplified in Figure 2. For questions asking about the cause of an event, we set the cause to the context and the effect to the candidate answer, while for questions asking about the effect, we reverse the direction.

Event	LM	Decoding	Outcome
The man turned on the faucet.	GPT	Top-k 10	the water started running in a steady stream
		Top-p 0.9	water began to flow out of the faucet and onto
		Beam 5	the water began to boil
	GPT-2 S	Top-k 10	his face became pale
		Top-p 0.9	the fire broke out
		Beam 5	he was able to get out of the car and
	GPT-2 XL	Top-k 10	the man's blood was sprayed everywhere, and his
		Top-p 0.9	the water in the bathtub began to flow
		Beam 5	water gushed out of the sink
The man received a parking ticket.	GPT	Top-k 10	the next day they were forced to drive around town
		Top-p 0.9	he was a bit confused about the situation, but
		Beam 5	he had to pay for the ticket
	GPT-2 S	Top-k 10	he has to pay the fare on the spot
		Top-p 0.9	he left his job
		Beam 5	he was arrested
	GPT-2 XL	Top-k 10	the ticket was sent to the city's Department of
		Top-p 0.9	he went to his car and pulled out a
		Beam 5	he was arrested and charged with violating the city's

Table 2: Example outcomes generated for COPA events (conditioned on "`[context]`. As a result,").

Following Shwartz et al. (2020b), we compute the cross entropy loss of each statement, and predict the candidate answer associated with the statement with the lowest loss (most plausible statement). Figure 2

(a) Accuracy score and the average rank of gold color.

LM	Pre-trained		Fine-tuned	
	Acc.	Rank	Acc.	Rank
Majority	35.8			
BERT	51.9	40.6	**69.2**	**1.7**
BERT-L	56.4	40.14	**70.1**	**1.68**
RoBERTa	49.0	63.4	**67.8**	**1.74**
RoBERTa-L	55.4	49.7	**68.7**	**1.71**

(b) Example sentences along with top 3 color predictions for each of the pre-trained models and the fine-tuned models (+FT). We note that the predictions are sensitive to phrasing.

Sentence	Majority	BERT-L	RoBERTa-L	BERT-L+FT	RoBERTa-L+FT
The __ banana is tasty.	g o y	b r w	b r be	g b r	g y r
The __ apple is sweet.	g r o	b w g	r b be	g r b	g r y
The __ cat is cute.	b w be	be b w	b gy r	p w b	b w o
The __ dove is beautiful.	w bn r	b be w	r w b	w be bn	w b be
The __ cow eats grass.	r w b	bn b be	r b be	b r b	bn r b
The __ dog runs in the park.	b w y	b bn w	w b be	b be y	b y w

Table 4: Performance and example predictions for the color prediction experiment. For each of BERT and RoBERTa, we report the performance of the pre-trained only model and the model fine-tuned on the color train set. The majority baseline predicts the most common color associated with the following noun in the train set, e.g. $\mathrm{majority}(\text{banana}) = \text{green}$.

illustrates the model's prediction for a given COPA instance.

Table 3 shows the accuracy on the development set across different LMs. The GPT models slightly improve upon the majority baseline.

3.1 Disconfirmed Expectations

G&V suggest that a better source for typical outcomes is textual constructions that indicate a speaker's expectation about the world was not met. For example, "Sally crashed her car into a tree but wasn't hurt" indicates that if a person crashed their car, they are likely to be hurt. An initial exploration of this approach was done by Gordon and Schubert (2011), but they concluded that extracting this type of rules from corpora is limited due to the sparseness of the clauses and the discourse patterns.

We conjecture that neural LMs may overcome the sparseness issue and be used for both scoring and generating typical outcomes. We therefore extend the zero-shot model by adding *disconfirmed expectations* (Zero-shot+DE) to the supporting statements: `[cause] [negative discourse marker] ([surprise expression]) [negated effect]`. We recognize the main verb of the effect using SpaCy and negate it to create the negated effect statement.

Table 3: Accuracy on the COPA development set.

LM	Zero-shot	Zero-shot+DE
Majority	0.55	0.55
GPT	**0.59**	0.56
GPT2-S	0.58	**0.59**
GPT2-XL	**0.61**	0.60
XLNet-S	**0.55**	0.49
XLNet-L	**0.43**	0.42

The results in Table 3 show that adding disconfirmed expectations usually degrades the performance. We observed that this often happens when a statement of the form "`[context] [negative discourse marker] [negated wrong answer]`" is incorrectly ranked as plausible, as in "He ran out of onions. Yet, for some reason the cook's eyes did not water". While the LM recognizes the lexical relatedness between onions and watering eyes, it is not sensitive to negation, as was recently shown for several other language models (Ettinger, 2020; Kassner and Schütze, 2020).

4 Properties

According to G&V, people are more likely to state unusual properties of a concept (*blue pencil*) than usual ones (*yellow pencil*). Recently, Weir et al. (2020) studied LMs' ability to associate concepts with their properties, by providing the LM the concept and predicting the properties and vice versa. Overall, LMs performed reasonably well, with RoBERTa outperforming BERT. Both performed better on encyclopedic and functional properties ("A bear is an animal") than on perceptual properties, which are less often mentioned in text (Collell Talleda and Moens, 2016; Forbes et al., 2019).

We hypothesize that while LMs are to some extent capable of learning association between concepts and their properties indirectly by aggregating across contexts, during this process, they often overgeneralize, predicting semantically-similar but mutually exclusive values. We verify that by evaluating

BERT and RoBERTa's ability to predict colors. We constructed a list of 11 common colors and extracted all sentences in Wikipedia in which a color modifies a noun, masking the color tokens (e.g. "A bear is [MASK]"). We then split the data into train (1,169,590 sentences) and test (10,000 sentences).

Table 4a presents the results of pre-trained-only LMs vs. LMs fine-tuned on the train set, with a masked LM objective, to predict the color. First, we note that the pre-trained BERT models outperform the RoBERTa model, which is expected given that BERT was already exposed to the sentences in the dataset during pre-training on Wikipedia. Despite that, the fine-tuned models still exhibit a dramatic boost in performance, both in terms of accuracy and average rank of the correct color. This is an encouraging result: it's possible to correct the over-generalization by further exposing the LM to the "truth". With that said, this corpus-based "truth" is not a ground truth, and given that the sentences were not manually verified, it is still biased towards the unusual, containing strange concepts like "blue cat".

5 Related Work

Commonsense in pre-trained LMs. There is ongoing research on extracting commonsense knowledge from pre-trained LMs, providing mixed results. On the one hand, Petroni et al. (2019) and Davison et al. (2019) somewhat successfully used pre-trained LMs to complete commonsense KBs. On the other hand, Logan et al. (2019) have shown that LMs are limited in their ability to generate accurate factual knowledge, and Kassner and Schütze (2020) and Ettinger (2020) pointed out that LMs are not sensitive to negation, resulting in generating incorrect facts ("birds can't fly"). Finally, Shwartz et al. (2020b) showed that despite being noisy, knowledge generated by LMs can be used to improve performance on commonsense tasks.

Similarly to our color experiment, Bouraoui et al. (2020) developed a LM-based relation classification model that included a color relationship. The model starts with a seed of known word pairs for a given relationship, uses it to find template sentences indicative of the relationship, and fine-tunes BERT on these retrieved sentences. Their experiment had a different purpose from ours, in which we probed the LMs for knowledge already captured by their pre-training phase.

Learning from other modalities. Much of our world knowledge is innate or acquired through modalities such as vision, including physical commonsense ("physical objects can't be in different places at the same time") and social commonsense ("people do and say things for reasons"). There has been little work on learning meaning from other modalities (Kiela and Clark, 2015; Zellers et al., 2019), but there is a shared understanding in the community that this is the imperative next step (Bisk et al., 2020; Bender and Koller, 2020).

6 Conclusion

We show that pre-trained LMs to some extent overcome reporting bias in the sense that they possess knowledge that wasn't explicitly stated, including trivial facts. Unfortunately, they also over-represent rare and newsworthy events, amplifying the bias that already exists in their training corpus.

The results in this paper are in line with prior work that showed that LMs amplify social bias (May et al., 2019; Sheng et al., 2019) and knowledge about named entities that are prominent in the corpus (Shwartz et al., 2020a). Going forward, it is important to study how the choice of training corpus, model size, and other factors affect the type and extent of biases the LM would have.

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

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *EMNLP*.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *AAAI*.

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1.

Guillem Collell Talleda and Marie-Francine Moens. 2016. Is an image worth more than a thousand words? on the fine-grain semantic differences between visual and linguistic representations. In *Proceedings of the 26th International Conference on Computational Linguistics*, pages 2807–2817. ACL.

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, November. 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, June. Association for Computational Linguistics.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8(0):34–48.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? In *CogSci*.

Jonathan Gordon and Lenhart K Schubert. 2011. Discovering commonsense entailment rules implicit in sentences. *EMNLP 2011*, page 59.

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. ACM.

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, 7-8 June. Association for Computational Linguistics.

H Paul Grice, Peter Cole, Jerry Morgan, et al. 1975. Logic and conversation. *1975*, pages 41–58.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1).

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. Association for Computational Linguistics.

Douwe Kiela and Stephen Clark. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal, September. Association for Computational Linguistics.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

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.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Tom Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Bishan Yang, Justin Betteridge, Andrew Carlson, Bhavana Dalvi, Matt Gardner, Bryan Kisiel, et al. 2018. Never-ending learning. *Communications of the ACM*, 61(5):103–115.

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, November. Association for Computational Linguistics.

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. -.

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

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3398–3403, Hong Kong, China, November. Association for Computational Linguistics.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020a. "you are grounded!": Latent name artifacts in pre-trained language models. In *EMNLP*.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020b. Unsupervised commonsense question answering with self-talk. In *EMNLP*.

Mohammad S Sorower, Janardhan R Doppa, Walker Orr, Prasad Tadepalli, Thomas G Dietterich, and Xiaoli Z Fern. 2011. Inverting grice's maxims to learn rules from natural language extractions. In *Advances in neural information processing systems*, pages 1053–1061.

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

Arjen van Dalen. 2012. Structural bias in cross-national perspective: How political systems and journalism cultures influence government dominance in the news. *The International Journal of Press/Politics*, 17(1):32–55.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

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, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

A Action and Events

Action	Actual Frequency for Lifetime (Source)	Normalized Frequency				
		Actual	Corpus	BERT	RoBERTa	GPT-2
thinking	1,433,355,000 (50,000 per day)	5.26e-01	9.21e-02	1.74e-01	8.66e-03	5.74e-03
breathing	660,489,984 (23,040 per day)	2.42e-01	3.51e-03	2.04e-02	8.11e-03	2.89e-04
blinking	344,005,200 (12,000 per day)	1.26e-01	6.84e-04	1.63e-03	0	0
eating	86001.3: 3 times per day	3.16e-05	1.23e-02	2.64e-02	1.09e-02	1.45e-03
sleeping	28667.1: 1 time per day	1.05e-05	1.03e-02	1.19e-02	2.65e-02	6.33e-04
working	20420.4: 5 times a week	7.49e-06	5.66e-02	5.81e-02	7.59e-02	4.22e-03
exercising	8168.16: 2-3 times a week	3.00e-06	2.44e-02	0.00e+00	1.17e-03	2.14e-04
getting married	1.66: 0-3 times per life	6.09e-10	4.76e-03	5.37e-02	2.26e-01	6.40e-04
getting divorced	1: 0-2 times per life	4.04e-10	8.95e-04	6.91e-03	1.49e-02	3.72e-05
being born	1	4.04e-10	7.35e-02	7.76e-02	1.75e-02	4.55e-03
being named	1	4.04e-10	2.49e-01	1.07e-01	1.02e-02	3.44e-03
dying	1	4.04e-10	1.55e-01	3.72e-02	1.66e-01	1.39e-02
being abused	0.5 (source)	1.84e-10	7.43e-03	3.28e-02	2.83e-02	4.30e-04
being injured	0.1263 (Episodes per 1,000 population: 126.3)	4.64e-11	6.74e-02	6.94e-03	1.01e-01	6.45e-04
being raped	0.01 (18.3% of women (50.8% of population) and 1.4% of men (49.2% of population))	3.66e-11	3.51e-04	1.03e-02	3.59e-02	1.06e-04
being killed	4.01×10^{-2} (murder + 1 out 28 in accident)	1.47e-11	2.59e-02	4.57e-02	3.32e-02	1.19e-03
being arrested	0.031526 (3,152.6 arrests per 100,000)	1.16e-11	5.06e-02	5.23e-03	9.85e-02	2.52e-03
being adopted	0.021 (7 million out of 328.2)	7.83e-12	4.93e-03	4.54e-03	8.53e-03	3.24e-05
being murdered	4.37×10^{-3} (1 in 229 deaths)	1.60e-12	2.99e-02	5.15e-02	7.88e-02	1.34e-03
being abandoned	0.000175 (7000 each year, out of 4M births)	6.42e-14	6.45e-04	4.17e-03	1.15e-02	3.46e-05

Table 5: Frequency of actions performed or occurring to a person during their lifetime, along with the sources used for actual frequency calculation, and the normalized scores for actual frequency, corpus (Google Ngrams), and LM scores. Daily statistics were multiplied by 365 × 78.54 (average life expectancy in the US: `https://www.cdc.gov/nchs/fastats/life-expectancy.htm`).

Action	Action Terms
thinking	thinking, thinks, think, thought
breathing	breathing, breathe, exhale, inhale
blinking	blinking, blink, blinks, blinked
talking	talking, talk, talked, say, said, saying, converse, conversed, conversing
eating	eat, eating, ate, dine, dining, dined
sleeping	sleeping, sleep, sleeps, slept
working	working, work, worked, employed
exercising	exercising, exercise, exercised
getting married	married
getting divorced	divorced
being born	born
being named	named, called
dying	died, die, dies, dying
being injured	injured
being arrested	arrested
being murdered	murdered, killed
being killed	killed
being raped	raped
being abused	abused, molested, assaulted, beat, bullied, oppressed, tortured
being shot	shot
being adopted	adopted
being abandoned	abandoned

Table 6: Synonyms and subactions used for each action in Section 2.