EECS 595 Final Report: Evaluating XLNet on Trip Dataset

Le Qin leqin@umich.edu

Abstract

In recent years, large pre-trained language models, from GPT-3 to Bert, has draw broaded attention in NLP fields, and inspired tons of related inspections for years. With more datasets and higher number of parameters, they outperform humans in many NLP subtasks, and are achieving higher scores each year. However, concerns are also raised from these seemingly perfect scores. The interprebility of machine learning models have always been a problem, and researcheres are curious about whether machines truly understand the reasoning behind these NLP tasks, or just learn the superficial labeling skills. More recently, a new commonsense dataset TRIP, (Tiered Reason- ing for Intuitive Physics), are brought up to deal with these concerns, with three valuable key metrics: accuracy, consistency, and verifiablilty. By evaluating this dataset on large machine models, we are hoping to reveal hints on whether they can do the real reasonings. In this paper, we examined the performance of XLNet on TRIP, and

1 Introduction

In recent years, lots of benchmarks are developed from large-scale pre-trained language models and are proved to be effective for many natural language processing problems, such as BERT, (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), RoBERTa, (Robustly Optimized BERT)(Liu et al., 2019), DeBERTA, (Decoding-enhanced BERT) (He et al., 2021). All these three models are variant of Transformers (Vaswani et al., 2017), which apply self-attention in computation, and achieve greater results than recurrent neural networks (RNN). Nowadays, these fine-tuning based models can achieve excellent performance on sentence-level as well as token-level tasks, such as commonsense reasoning, (Talmor et al., 2019), commonsense inference, (Bowman et al., 2015) and much more. These wonderful results also stimulate people towards using more data to train models, and resulted in even larger amount of parameters. For example, in the first generation of GPT, the total number of parameters are 117 Million, (Radford et al., 2018), which is about the same number of Bert-Base 110 Million, (Devlin et al., 2019) at that time. However, two years later, GPT-3 has more than 175 Billion parameters (Brown et al., 2020). It's no doubt that higher number of parameters brought better performance, yet whether they have a truly understanding about their tasks remained in mystery. Previous works from Gururangan, et al. mentioned that, for text classification task, some specific linguistic phenomena like negation can be highly correlated to some label classes, resulted in machine models are doing the inference tasks without the needs to understand whole text. (Gururangan et al., 2018). Moreover, Poliak et al suggests that statistical irregularities may reduce the difficulties of NLP tasks into natural language inference, which allow a model to do classification without accessing to the text content. As a result, to address these concerns, Storks, et al brought up a Tiered Reasoning for Intuitive Physics (TRIP) dataset (Storks et al., 2021). This dataset consists of pairs of stories, which one of them is plausible, and anothere is not. Every story is of short sentences, and the task for large models is to determine which one of them is plausible, and the break point where story cannot be achieved in physic worlds, and also the physic annotations changed along it.

In this paper, we will mainly focus on the performance of XLNeet on TRIP, like in the origonal TRIP paper, XLNet will need to determine the plausible story, the pair of conflicing sentences in the implausible story, and which of physical states leads to the conflicting story.

The contributions of this work can be it tried to implement on other large pre trained models other than Bert, Roberta, and Deberta. The results can

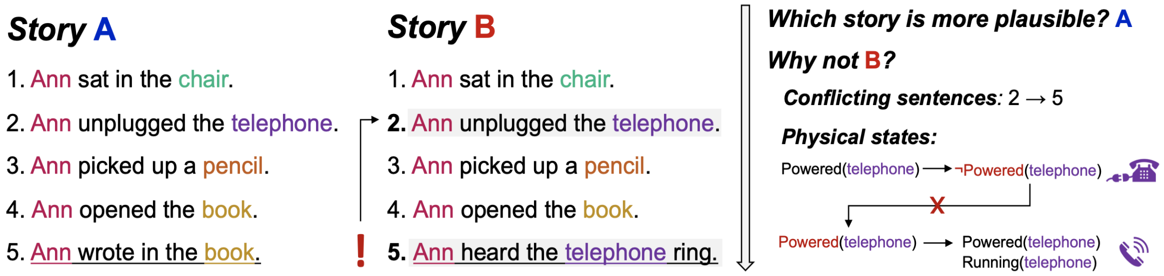

Story A
1. Ann sat in the chair.
2. Ann unplugged the telephone.
3. Ann picked up a pencil.
4. Ann opened the book.
5. Ann wrote in the book.
Story B
1. Ann sat in the chair.
2. Ann unplugged the telephone.
3. Ann picked up a pencil.
4. Ann opened the book.
5. Ann heard the telephone ring.
Which story is more plausible? A
Why not B?
Conflicting sentences: 2 → 5
Physical states:
Powered(telephone) → ¬Powered(telephone)
Powered(telephone) → Powered(telephone) Running(telephone)

Figure 1: A story pair from TRIP dataset (Storks et al., 2021)

be used to support that reasoning process might be lost in many current employed large machine models, and motivating future work on solve these problems.

2 Related Work

2.1 Tiered Reasoning for Intuitive Physics

Using artificial intelligence to solve physics commonsense reasoning calls people's attention on a new field that are making slow progress a few years ago: ordinary commonsense (Davis and Marcus, 2015). It focus on problems that can only be solved using a certain amount understanding about the real world. Later, many people are drawn to these subfields (Marcus, 2018; Sap et al., 2019; Lin et al., 2019). For humans, it seems that children learn common sense reasoning as they grow up, and benefited with daily interaction with environment around them (Bliss, 2008). They then argue that human reasonings are consists of different schemes, where schemes are interacting with each other. By contrast, machines seem to struggle with reasoning. Evidence shown that although large neural networks can related objects with tons of related information, it fails to capture more subtle interplay properties (Forbes et al., 2019).

2.2 TRIP dataset

TRIP dataset is developed to facilitate discoveries in this domain. Stories in this dataset are all written by human authors, with every story is of concrete physical actions (Storks et al., 2021). Every two stories are connected as a pair, which only one difference in it is that one of them consists a sentence that will make the whole story implausible. Moreover, each story are consists dense physical annotation words. An example from Storks et al are shown in figure above. Line 5 is the difference in these two stories, which makes Story B implausible. Then, the line 2 is the breakpoint, as the telephone cannot ring after the telephone is unplugged. The task for machine models then is to first determine which story is possible to happen in real world. For such predictions to be made, one must have the knowledge of verb causality, the ability to sense the change of states of an object from the verb world, such as after melting, the object will be in liquid form; It have to know precondition, that for an object to be cut, it has to be in solid form; It also have to know the rules of intuitive physics, for example, two solid objects cannot pass each other (Storks et al., 2021). To minimize objectivity, each author are asked to write simply and declarative sentences related to concrete actions that can be visualized in the physical world (Storks et al., 2021). To maintain the plausibility in longer context, unlike previous work that only have one sentence of context (Zellers et al., 2018), authors are asked to write at least five sentences long stories, and each sentence of them should be plausible solely (Storks et al., 2021). In such way, the story can be less influenced by distributional biases. Moreover, three levels of 20 physical annotations are provided in this dataset to enable systematic review, which corresponds to the three tasks.

2.3 Tiered Baseline for TRIP

Since our project is mainly developed on Storks et al's paper (Storks et al., 2021), we adapted to their baseline structures.

Contextual Embedding

This module is implemented with a pre-trained language model. In our case, we used HuggingFace pre-trained XLNet as our model (Wolf et al., 2020). Our input stays the same with Storks et al's work, which takes an input sentence, the name of entity, and an entity-first input formulation. The output then is a contextualized numerical representation of it (Storks et al., 2021).

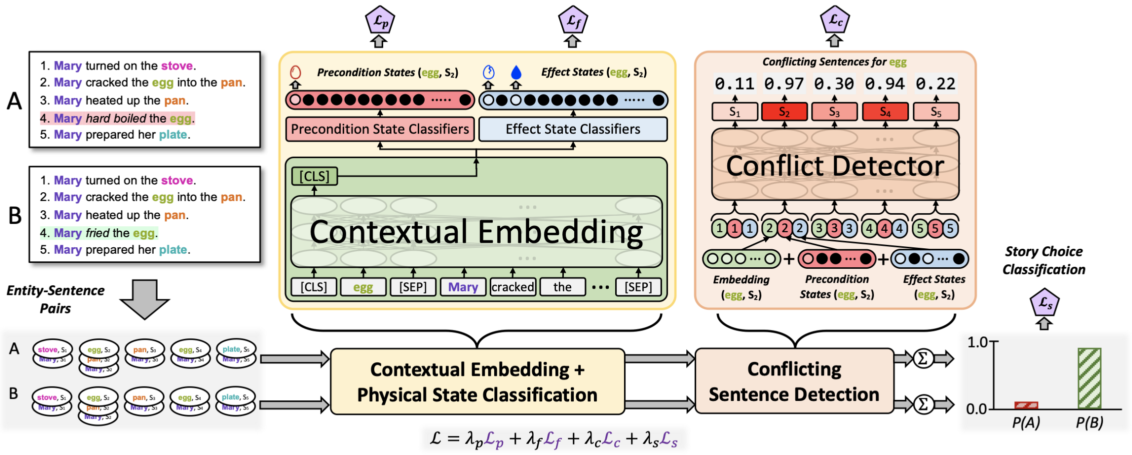

A
1. Mary turned on the stove.
2. Mary cracked the egg into the pan.
3. Mary heated up the pan.
4. Mary hard boiled the egg.
5. Mary prepared her plate.
B
1. Mary turned on the stove.
2. Mary cracked the egg into the pan.
3. Mary heated up the pan.
4. Mary fried the egg.
5. Mary prepared her plate.
Entity-Sentence Pairs
Precondition States (egg, S₂)
Effect States (egg, S₂)
Precondition State Classifiers
Effect State Classifiers
[CLS]
Contextual Embedding
[CLS] egg [SEP] Mary cracked the ... [SEP]
Conflicting Sentences for egg
0.11 0.97 0.30 0.94 0.22
S₁ S₂ S₃ S₄ S₅
Conflict Detector
Embedding (egg, S₂)
Precondition States (egg, S₂)
Effect States (egg, S₂)
Story Choice Classification
Contextual Embedding + Physical State Classification
Conflicting Sentence Detection
Σ
1.0
0.0
P(A)
P(B)
$\mathcal{L}_p$
$\mathcal{L}_f$
$\mathcal{L}_c$
$\mathcal{L}_s$
$\mathcal{L} = \lambda_p\mathcal{L}_p + \lambda_f\mathcal{L}_f + \lambda_c\mathcal{L}_c + \lambda_s\mathcal{L}_s$

Figure 2: Tiered reasoning system structure (Storks et al., 2021)

Precondition and Effect Classifiers

In this module, we have one precondition classifier and one effect classifier for each of the 20 physical attributes. Softmax is then used for generating the output (Storks et al., 2021).

Conflict Detector

The task of this module is to predict whether there is conflicts existing in the entity's physical state, and find a pair of sentences that might be the cause of it. Another transformer is used at here, but the input is the contextual embedding, and the classification logits. The output then is the probability of each sentence conflicting with another sentence in the story (Storks et al., 2021).

Story Choice Prediction

The last remaining task is then to output which story is classified as plausible. Given the output from last module, we sum negative outputs and apply softmax for output (Storks et al., 2021).

2.4 BERT

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019), is built on Transformer networks (Vaswani et al., 2017) to predict a masked word from its context, and to classify whether two sentences are consequent to each other or not.

Opposed to directional models that read the text input sequentially, Transformer encoder in Bert reads the entire sequence of words at once. By adding this feature, model can learn the context of a word base on its surroundings. Masked words from input are replaced with Masked LM(MLM), with 15% of the words in each sequence to be replaced by a [MASK] token.

During training, Next Sentence Prediction is used, that 50% of the inputs are a pair of sequential sentence in the original document, and the rest 50% of the sentence is choose randomly in the corpus.

The goal in the whole training process is to minimize the combined loss of these two strategies.

2.5 DeBERTa

DeBERTa, (He et al., 2021), is also a Transformer based neural networks (Vaswani et al., 2017). The motivations for this task is that the attention for a sentence should not only depend on the words in it, but also their relative positions. For example, dependency of adjacent words will be stronger than split words. Therefore words in the input layer and the positions of word in sentences are reconsidered in DeBERTa (Vaswani et al., 2017). Unlike words are represented using a vector of sum of their content embedding and position embedding, words are represented using two vectors that take care of content and position respectively (He et al., 2021). In this manner, disentangled attention is used to represent the strength of position as well as content. The enhanced mask decoder also take positions of words in sentences into consideration. Moreover, unlike relative positions used in BERT, DeBERTa used the absolute position in modelling processes, allowing syntactical nuances to play their roles (Vaswani et al., 2017).

2.6 RoBERTa

RoBERTa, Robustly Optimized BERT (Liu et al., 2019). Compared to BERT, it made several ad-

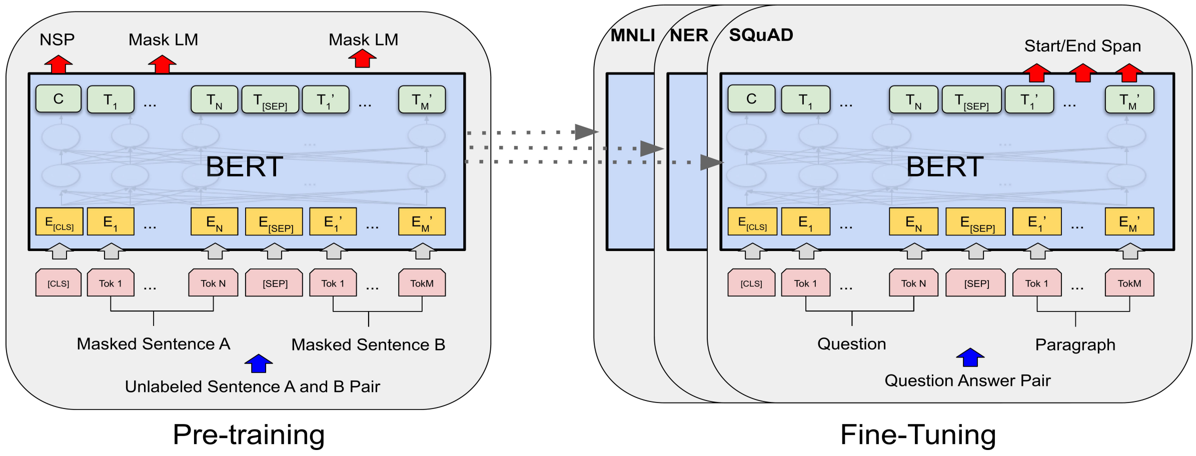


Figure 3: BERT Architecture (Devlin et al., 2019)

justments: 1) It used larger batch size (2k and 8k), compared with original 256 in BERT (Liu et al., 2019) and therefore longer time in training. 2) It used dynamic masking, that mask is done when feeding the input to the model, compared to BERT that perform masking in data preprocessing (Liu et al., 2019). In this way, same mask can be avoided using on training instance in every epoch.

2.7 XLNet

In BERT, [MASK] are used in pre-training, while in fine-tuning [MASK] is unavailable. Therefore, a disagreement exist in these two modes. To address this issue, XLNet used auto regressive LM, with the sentence still take inputs from left to right, but have both the context_before and context_ after (Yang et al., 2019). To achieve that, XLNet used Permutation Language Model, which will randomly do permutations for a given sentence, with a word fixed. Then, they will also be used as input for LM. In this way, for the fixed word x, its before_context and after_context can all be used during training, while stays in the form of predicting a word from left to right (Yang et al., 2019).

3 Approaches

We mainly use XLNet from Huggingface as our pre-trained model (Wolf et al., 2020). Then we perform grid search to find the best hyperparameters combinations. The setting of training and evaluations are same as Storks et al's work in order to facilitate comparison in section 4.

4 Evaluations

4.1 Evaluation Metrics (Storks et al., 2021)

The following metrics are used in order to measure machines' ability in reasoning task.

Accuracy. (Storks et al., 2021)

The proportion of plausible stories are correctly identified.

Consistency. (Storks et al., 2021)

The proportion of both plausble stories as well as conflicting pairs of stories are correctly identified.

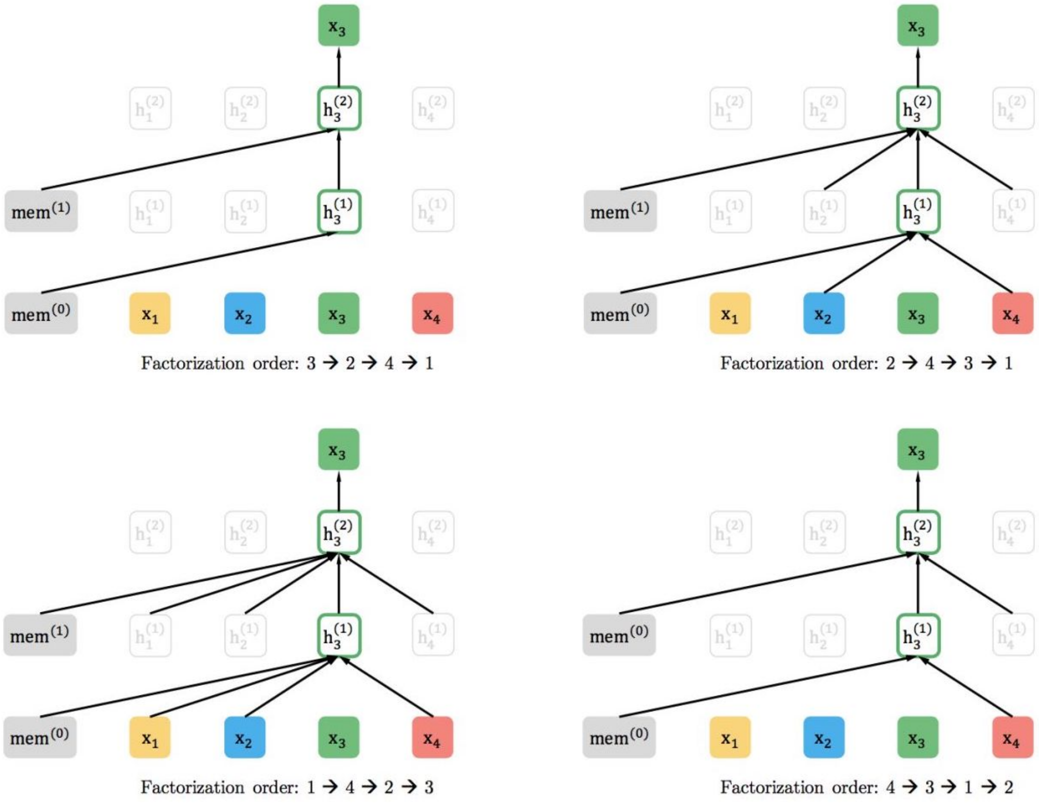
Verifiability. (Storks et al., 2021)

The proportion of stories that not only fulfill consistency, but also identify the underlying changed physical states. If we suppose accuracy a, consistency b, verifiability c, then for a reliably coherent machine model, $a \approx b \approx c$ (Storks et al., 2021).

4.2 Results

Four loss functions are used during the training: $\mathcal{L}_p$ for precondition classification, $\mathcal{L}_f$ for effect classification, $\mathcal{L}_c$ for conflicting sentence detection, and $\mathcal{L}_s$ for story choice classification (Storks et al., 2021). The results of using these loss functions on XLNet as well as previous benchmarks are shown in Table 1. We can notice that, its performance is similar to previous work, that achieve high on accuracy, but stay lower on consistency and verfiability.

When fine-tuning RoBERTa's contextual embedding directly to the end task, it achieved up to 97% accuracy, but then have pretty low verifiability (Storks et al., 2021). Our model also showed this pattern along training processes.



Factorization order: 3 → 2 → 4 → 1

Factorization order: 2 → 4 → 3 → 1

Factorization order: 1 → 4 → 2 → 3

Factorization order: 4 → 3 → 1 → 2

Figure 4: Illustration of permutation language modeling (Yang et al., 2019).

4.3 Discussion

My group got questions about is there any reason to use XLNet. We choose to use XLNet came from a very intuitive thought, but it might be fun to think about are there any structures of XLNet that might correspond to the results in my experiment. The permutation modeling structure that enable both above and below context to be referenced might be a reason for it. If we have more time, it might worthy experimenting with.

4.4 Conclusion

In this project, we used TRIP, a dataset for physical commonsense reasoning on XLNet. Several variations are used. Our results shows that XLNet, although perform well on classification tasks, fails to stay consistent and verifiable for underlying physical reasoning.

4.5 Github Repository Link

The table related NoirChad

References

Joan Bliss. 2008. Commonsense reasoning about the physical world. *Studies in Science Education*, 44(2):123–155.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *arXiv preprint arXiv:1908.02899*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith.

Model	Accuracy (%)	Consistency (%)	Verifiability (%)
random	47.8%	11.3%	0.0%
All losses			
BERT	**78.3**	2.8	0.0
ROBERTA	75.2	6.8	0.9
DEBERTA	74.8	2.2	0.0
XLNet	76.7	12.7	0.0
Omit story choice loss $\mathcal{L}_s$			
BERT	**78.3**	2.8	0.0
ROBERTA	75.2	6.8	0.9
DEBERTA	74.8	2.2	0.0
XLNet	77.5	21.37	4.3
Omit Conflict Detection Loss $\mathcal{L}_c$			
BERT	73.9	**28.0**	9.0
ROBERTA	73.6	22.4	**10.6**
DEBERTA	75.8	24.8	7.5
XLNet	77.5	21.37	4.3
Omit State Classification Losses $\mathcal{L}_p$ *and* $\mathcal{L}_f$			
BERT	75.2	17.4	0.0
RoBERTA	71.4	2.5	0.0
DEBERTA	72.4	9.6	0.0
XLNet	80	0	0

Table 1: Tiered classfiers on the validation set of TRIP. Random baseline BERT, RoBERTa and DeBERTa are came from Storks et al, including for future comparisons. (Storks et al., 2021).

2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*.

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.

Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

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

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.

Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. 2021. Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.