# Do Transformers Dream of Inference, or Can Pretrained Generative Models Learn Implicit Inferential Rules?

**Zhengzhong Liang and Mihai Surdeanu**
Computer Science Department, The University of Arizona
1040 4th St, Tucson, AZ 85721, USA
{zhengzhongliang, msurdeanu}@email.arizona.edu

## Abstract

Large pretrained language models (LM) have been used successfully for multi-hop question answering. However, most of these directions are not interpretable, as they do not make the inference hops necessary to explain a candidate answer explicitly. In this work, we investigate the capability of a state-of-the-art transformer LM to generate explicit inference hops, i.e., to infer a new statement necessary to answer a question given some premise input statements. Our analysis shows that such LMs can generate new statements for some simple inference types, but performance remains poor for complex, real-world inference types such as those that require monotonicity, composition, and commonsense knowledge.

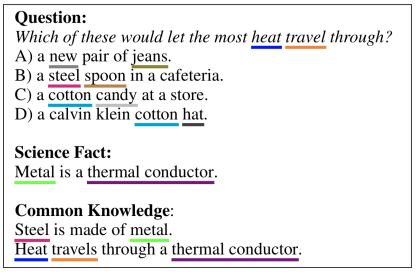


Figure 1: An example of question and candidate answers from OpenbookQA (Mihaylov et al., 2018) (the correct answer is option B). The science fact and the commonsense knowledge facts are needed to explain the correct answer. Usually the large LMs solve this problem by taking the question, the science fact, the common knowledge facts and each candidate answer as the input and producing a single score indicating the probability of the candidate answer being justified by all of the inputs. But *why* the facts explain the answer is normally not covered.

## 1 Introduction

The emergence of large pretrained language models (LM) (Devlin et al., 2019; Liu et al., 2019) yielded significant progress in question answering (QA), including complex QA tasks that require multi-hop reasoning (Banerjee et al., 2019; Asai et al., 2019; Yadav et al., 2019). Most of these state-of-the-art (SOTA) approaches address multi-hop reasoning tasks in a discriminative manner: they take the question, the candidate answer, and all the context available as the input, and produce a single score indicating the likelihood of the answer as justified by the provided context (an example is shown in Figure 1). However, *why* that context actually justifies the answer remains unclear to the human end user of the QA system.

In contrast, most of us are likely to answer the question in Figure 1 by building a reasoning chain from the given facts. For example, such a possible chain starts by first combining "metal is a thermal conductor" and "steel is made of metal' to yield "steel is a thermal conductor". Next, combining "steel is a thermal conductor" and "heat travels through a thermal conductor" yields "heat travels through steel". And, finally, "heat travels through steel" supports the correct explanation that "a steel spoon in a cafeteria would let the most heat travel through." Generating such reasoning chains can be crucial for the adoption of natural language processing applications such as QA in critical domains such as medical or law.

Motivated by this, in this work we investigate whether a state-of-the-art (SOTA) transformer-based language model is able to generate a valid intermediate statement given two premise statements on a natural language QA dataset, which is fundamental to generating the reasoning chains. Our results show that although the SOTA model investigated can handle some types of inferences well, there remain multiple types of inferences where the LM fails.[1]

[1]The code and data for our analysis can be found at https://github.com/clulab/releases/tree/master/emnlp2020-generative-nli.



*Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 76–81
Online, November 19, 2020. 
ISBN 978-1-952148-66-8

| Category | Without Hint | With Hint |
|---|---|---|
| Perfect | 31/87 | 43/87 |
| Acceptable | 11/87 | 13/87 |
| Unacceptable | 45/87 | 31/87 |

Table 1: Statistics of the quality of the generated T5 statements on the dev set of QASC. The same randomly sampled 87 examples are manually evaluated for their quality, in both the "without hint" and "with hint" configurations.

## 2 Related Work

Recently several works have investigated whether deep learning (DL) language models (LM) are able to learn and use the explicit and implicit rules in natural language. (Sinha et al., 2019) build a synthetic dataset containing the relationships between people; their language model needs to predict the unstated relationships between people. The problem can be summarized as: given that "Mike is the child of Kate and Kate is the child of Tom", the model needs to predict "Tom is the *grandparent* of Mike", by learning the implicit rule: "If X is the child of Y and Y is the child of Z, then Z is the grandparent of X". It has been shown that the transformer networks perform well on this task.

Other works have analyzed whether DL language models are able to leverage explicit rules. (Clark et al., 2020) generates a synthetic dataset consisting of facts and rules. The problems can be summarized as: given the facts such as "X is red" and "X is big", as well as rules such as "If X is red and big, then X is strong", the LM trained on this data must be able to judge whether "X is strong" is true. They demonstrate that transformers can fulfill this task well, and are able to generalize to unseen lexicons.

However, all existing works investigate this problem in a discriminative manner: either a single score, a single token, or a single choice is produced as the output. In contrast, we conduct our work in a generative manner: the LM needs to generate a whole natural language statement as the output. We believe this task will eventually give the LM the ability to generate clear and complete explanations, which are necessary in multi-hop reasoning problems. Further, we investigate the capability of transformers to generate inferential statements on a complex, real-world task in the science domain, which relies on much sparser data than other tasks previously investigated.

## 3 Approach

### 3.1 Problem Formulation

In this paper, we concentrate on a single-hop inference problem. That is, given the statements $S_1(A, B)$ and $S_2(B, C)$, the model needs to generate the valid and reasonable statement $S_3(A, C)$. Unlike reasoning tasks on structured knowledge bases or ConceptNet where $A$, $B$, $C$ are entities, here $A$, $B$ and $C$ can be any text in natural language: they can be words, phrases, or clauses.

We used the QASC dataset (Khot et al., 2020) for this task. QASC contains approximately $10,000$ questions in the science domain, where each answer is associated with two supporting facts (fact 1 and fact 2). These two supporting facts have tokens in common, which is necessary for our inference task that requires overlap between facts (through B). Importantly, for each answer QASC provides a *combined fact* that explains the answer, and which is directly inferred from the two supporting facts. The first two columns in Tables 2, 3, and 4 show a few examples of the supporting facts and the resulting combined fact. The forms of the combined facts can be very diverse due to the annotation process of QASC, where each annotator is first given fact 1, then the annotator needs to find an arbitrary fact 2 that has overlaps with the fact 1, and composes the combined fact, without other restrictions (Khot et al., 2020). [2] The task we investigate here is whether transformer-based LMs can infer the combined fact when provided with the two initial facts.

### 3.2 Method

We use the pre-trained Google T5 small model (Raffel et al., 2020) published by huggingface (Wolf et al., 2019), and fine-tune it on the QASC dataset.[3] We explore two types of input format:

**fact 1 + fact 2 → combined fact:** In this setting, T5 takes the two facts as input to generate the combined fact. The T5 input format is "substitution statement 1: [fact 1] statement 2: [fact 2]", where "substitution", "statement 1:" and "statement 2:" are user-defined keywords for the task.

[2] Note that the two supporting facts and the gold combined fact of each question in QASC are annotated by the creators of the QASC dataset, not by the authors of this paper.

[3] We used the Adam optimizer with a learning rate of 1e-4, as recommended in the tutorial. The training stops when the evaluation loss starts to increase; we allowed a maximum of 10 epochs of training.

| Input | Target | Prediction | Evaluation |
|---|---|---|---|
| *substitution statement 1:* if weather is stormy then there is a greater chance of rain. *statement 2:* rain is also known as precipitation. | if weather is stormy then there is a greater chance of precipitation. | if weather is stormy then there is greater chance of precipitation. | Perfect, exactly the same |
| *substitution statement 1:* lightning can cause a forest fire. *statement 2:* forest fires are examples of wildfires. | wildfires are caused by lightning. | lightning can cause wildfires. | Perfect, semantically the same |
| *substitution statement 1:* whiskers are used for feeling by animals. *statement 2:* whiskers whiskers are the stiff hairs on the dog's face. | stiff hairs on the face are used for feeling by animals. | stiff hairs on the dog's face are used for feeling by animals. | Acceptable |
| *substitution statement 1:* lymph is a fluid that leaks out of capillaries into spaces between cells. *statement 2:* bacteria invade lymph nodes, which swell, creating the buboes. | swelling nodes leak fluid out of capillaries into spaces between cells. | bacteria can swell. | Unacceptable, unreasonable or contradicting information |
| *substitution statement 1:* absorption of nutrients occurs mainly in the small intestine. *statement 2:* diarrhea reduces the absorption of all nutrients. | diarrhea reduces funtcion that occurs mainly in the small intestine. | diarrhea reduces the absorption of nutrients. | Unacceptable, missing essential elements |

Table 2: Output of T5 on QASC in the "without hint" configuration.

| Input | Target | Prediction | Evaluation |
|---|---|---|---|
| *substitution statement 1:* protozoa make humans sick when they become human parasites. *statement 2:* exhibit one-celled animals are called protozoa. | one-celled animals make humans sick when they become parasites. | one-celled animals make humans sick. | w/o hint |
| *substitution statement 1:* protozoa make humans sick when they become human parasites. *statement 2:* exhibit one-celled animals are called protozoa. *hint:* animals when become one-celled humans make parasites they sick. | one-celled animals make humans sick when they become parasites. | one-celled animals make humans sick when they become human parasites. | w/ hint, statement improved |
| *substitution statement 1:* mutualism is a symbiotic relationship in which both species benefit. *statement 2:* domestication of animals is an example of a symbiotic relationship. | domestication of animals is an example of mutualism. | domestication of animals is an example of mutualism. | w/o hint |
| *substitution statement 1:* mutualism is a symbiotic relationship in which both species benefit. *statement 2:* domestication of animals is an example of a symbiotic relationship. *hint:* is animals mutualism of domestication example an. | domestication of animals is an example of mutualism. | mutualism is an example of domestication of animals. | w/ hint, statement harmed |

Table 3: Comparison of T5 output in the "without hint" and "with hint" configurations on QASC.

**fact 1 + fact 2 + lexical hints → combined fact:** During our experiments, we noticed that sometimes multiple valid statements could be inferred from fact 1 and fact 2, which tended to confuse the LM.[4] To mitigate this issue, we added lexical hints to the model input, on what tokens would be best to be included in the generated statement. The terms in the hint are generated as $(\mathcal{Q} \cup \mathcal{A}) \cap (\mathcal{F}_1 \cup \mathcal{F}_2)$, where $\mathcal{Q}$ is the set of unique terms in the question, $\mathcal{A}$ is the set of unique terms in the answer, $\mathcal{F}_1$ and $\mathcal{F}_2$ are the sets of unique terms in fact 1 and fact 2.[5] This is inspired by the fact that each question in QASC is derived from the gold combined fact, so that even when multiple valid statements may be generated from fact 1 and fact 2, paying extra attention on the terms in the question and the correct answer is likely to force the model to make predictions related to the gold combined fact.

### 3.3 Evaluation Metric

For each configuration, we manually evaluated 100 generated statements against the corresponding gold combined fact on the dev set.[6] All generations are categorized into three classes.

**Perfect:** The generated statement is (1) exactly the same as the gold combined fact, or (2) semantically the same as the gold combined fact but uses a different expression.

[4] E.g., for the first and second row in Table 3, "one-celled animals make humans sick" is a valid generation, but not perfect w.r.t. the target.

[5] Thus, the text containing the lexical hints is simply a bag of words, rather than grammatical correct text.

[6] 13 data points had issues in the raw data, and were removed, leaving the actual number of data points analyzed as 87.

| Input | Target | Prediction | Question Type |
|---|---|---|---|
| *substitution statement 1:* skin color is a polygenic trait. *statement 2:* polygenic traits are the result of the interaction of several genes. *hint:* is genes of the result several skin color interaction. | skin color is the result of the interaction of several genes. | skin color is the result of the interaction of several genes. | Instantiation |
| *substitution statement 1:* if weather is stormy then there is a greater chance of rain. *statement 2:* rain is also known as precipitation. *hint:* stormy is greater weather there of a chance precipitation. | if weather is stormy then there is a greater chance of precipitation. | if weather is stormy then there is a greater chance of precipitation. | Equivalence |
| *substitution statement 1:* all cnidarians are aquatic. *statement 2:* cnidarians have a hydrostatic skeleton. *hint:* a are aquatic hydrostatic. | some aquatic animals have hydrostatic skeletons. | all aquatic animals have a hydrostatic skeleton. | Monotonicity with quantifier |
| *substitution statement 1:* absorption of nutrients occurs mainly in the small intestine. *statement 2:* diarrhea reduces the absorption of all nutrients. *hint:* occurs small mainly the diarrhea reduces in intestine . | diarrhea reduces function that occurs mainly in the small intestine. | diarrhea reduces the amount of food that occurs mainly in the small intestine. | Composition and summarization |
| *substitution statement 1:* kidney failure may be treated with dialysis. *statement 2:* kidney failure is a death sentence. *hint:* death dialysis. | a lack of dialysis may lead to death. | death can be treated with dialysis. | Need to rephrase to make the new statement reasonable |

Table 4: Output of T5 categorized by the types of the inference (w/ hint).

**Acceptable:** The generated statement is semantically valid, but its meaning is slightly different from the gold combined fact.

**Unacceptable:** The generated statement (1) contains contradicting information, or (2) has severe grammatically issues, or (3) is missing essential content from the gold combined fact (e.g., contains information from only fact 1 or only fact 2).

## 4 Results

Table 1 shows the overall statistics gathered by our analysis. All in all, our analysis shows that this inferential task is far from solved, with most of the inferred statements being not perfect. In particular, for the w/o hints configuration, less than half of the generated statements are perfect. Adding lexical hints to the input boosts the generation quality in general, but leaves 51% of inferences as not perfect. A detailed analysis of the generated statements highlights that T5 performs well in certain situations, and not so in others. We categorize below these situations, discuss some possible solutions, and leave a more systematic analysis of the reason why the model fails on some problems to a future study.

Below "well learned" means most of the predictions on that type of generations are evaluated as "perfect" and "not well learned" means most of the predictions are evaluated as "unacceptable" by the criteria mentioned in 3.3.

**Inference types well learned:**

**Instantiation** Here the input statements are $S_1(A, B)$ and $IsA(B, C)$, i.e., $C$ is an instantiation of a more general concept $B$. The target output is $S_1(A, C)$ (Table 4).

**Equivalence** Here the input statements are $S_1(A, B)$ and $Equ(B, C)$, i.e., $B$ is equivalent to $C$. The target output is $S_1(A, C)$ (Table 4).

**Inference types not well learned:**

**Multiple possible statements to generate** When the input statements are long and complex, there might be multiple valid statements that could be generated from the input (discussed in 3.2). In this case T5 tends to be confused. Adding lexical hints can relieve this problem to some extent by forcing the model to pay extra attention to certain areas in the input, but problems remain. First, even when adding the lexical hints, some generations are still not reasonable (Table 3). Second, accurately identifying the important fractions to pay attention to is itself a non-trivial problem. We believe this is an exciting area for future research. For example, some specialized architectures such as the pointer generator network (See et al., 2017) might be capable to learn what parts should be copied or ignored.

**Composition and summarization** As shown in the third to last row of Table 4, the new statement needs the composition of statement 1 and 2, and some summarization is needed (i.e., "absorption of nutrients" → "function").

**Dealing with quantifiers in natural language** As shown in the second to last row of Table 4, the new

statement needs complex monotonicity reasoning and the understanding of quantifiers.

**Generating statements that comply with commonsense knowledge** In several examples, the model generates statements that are grammatically correct but unreasonable regarding commonsense knowledge. In particular, many of these inferences require commonsense knowledge to generate new text and rephrasing to make the new statement reasonable. For example, in the last row of Table 4, "death can be treated with dialysis" is grammatically correct but unreasonable.

There might be multiple reasons why some types of generations are not well learned. For instance, it could be because the biases learned by T5 in the pre-training stage impede it from learning meaningful patterns by fine-tuning on a downstream task with relatively few training samples (e.g., the QASC dataset used in this paper has only about 8,000 training examples). Alternatively, it is possible that the patterns to be learned in this downstream task are too complex to be learned from the small training data available. We leave a more systematic analysis in this direction to future studies.

## 5 Conclusion

In this work we investigate how well a state-of-the-art transformer language model can generate a valid statement inferred from two given statements. We manually evaluated two fine-tuned T5 models (Raffel et al., 2020) with slightly different inputs (i.e., with and without contextual information) on the Question Answering via Sentence Composition dataset (Khot et al., 2020). Our analysis indicates that the two models can generate good-quality statements, when the inference relies solely on instantiation or equivalence. However, the models perform poorly on more complex inferences such as: (a) multiple valid statements can be generated given the premises, (b) inference that requires non-trivial reasoning of monotonicity (especially with quantifiers in natural language), (c) inference that needs composition and summarization, and (d) statements that require rephrasing based on background commonsense knowledge.

## References


Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2019. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations*.

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2019. Careful selection of knowledge to solve open book question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6120–6129, Florence, Italy. Association for Computational Linguistics.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *arXiv preprint arXiv:2002.05867*.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *AAAI*, pages 8082–8090.

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*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

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.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR:

A diagnostic benchmark for inductive reasoning from text. 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 4506–4515, Hong Kong, China. 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.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. Quick and (not so) dirty: Unsupervised selection of justification sentences for multi-hop question answering. 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 2578–2589, Hong Kong, China. Association for Computational Linguistics.