Is a Question Decomposition Unit All We Need?

Pruthvi Patel* **Swaroop Mishra*** **Mihir Parmar** **Chitta Baral**

Arizona State University

Abstract

Large Language Models (LMs) have achieved state-of-the-art performance on many Natural Language Processing (NLP) benchmarks. With the growing number of new benchmarks, we build bigger and more complex LMs. However, building new LMs may not be an ideal option owing to the cost, time and environmental impact associated with it. We explore an alternative route: can we modify data by expressing it in terms of the model's strengths, so that a question becomes easier for models to answer? We investigate if humans can decompose a hard question into a set of simpler questions that are relatively easier for models to solve. We analyze a range of datasets involving various forms of reasoning and find that it is indeed possible to significantly improve model performance (24% for GPT3 and 29% for RoBERTa-SQuAD along with a symbolic calculator) via decomposition. Our approach provides a viable option to involve people in NLP research in a meaningful way. Our findings indicate that Human-in-the-loop Question Decomposition (HQD) can potentially provide an alternate path to building large LMs[1].

1 Introduction

With the advent of large LMs, we have achieved state-of-the-art performance on many NLP benchmarks (Radford et al., 2019; Brown et al., 2020; Sanh et al., 2021a). Our benchmarks are evolving and becoming harder over time. To solve new benchmarks, we have been designing more complex and bigger LMs at the cost of computational resources, time and its negative impact on the environment. Building newer LMs for solving new benchmarks may not be an ideal and sustainable option over time. Inspired by humans, who often view new tasks as a combination of existing tasks, we explore if we can mimic humans and help the model solve a new task by decomposing (Mishra et al., 2021a) it as a combination of tasks that the model excels at and already knows.

As NLP applications are increasingly more and more popular among people in their daily activities, it is essential to develop methods that involve humans in NLP-powered applications in meaningful ways. Our approach attempts to fill this gap in LMs by providing a human-centric approach to modifying data. Solving complex QA tasks such as multi-hop QA, and numerical reasoning has been a challenge for models. Question Decomposition (QD) has recently been explored to empower models to solve these tasks with the added advantage of interpretability. However, previous studies on QD are limited to some specific datasets (Khot et al., 2020b) such as DROP (Dua et al., 2019) and HotpotQA (Yang et al., 2018). We analyze a range of datasets involving various forms of reasoning to investigate if "*a Question Decomposition Unit All We Need?*"

Figure 1 shows the schematic representation of a QD unit. The original question is difficult for a model to answer. However, it becomes easier for the model when a human decomposes the question into a set of simpler questions.

We manually decompose randomly selected 50 samples of each dataset. The decompositions we perform are purely based on intuitions to reduce the complexity of the question, inspired by the success of task-level instruction decomposition (Mishra et al., 2021a) in improving model performance. We experiment with GPT3 (Brown et al., 2020) and RoBERTA (Liu et al., 2019) fine-tuned on SQuAD 2.0 (Rajpurkar et al., 2018) and find that HQD significantly improves model performance (24% for GPT-3 and 29% for RoBERTa-SQuAD along with a symbolic calculator). Here, the evaluation happens on unseen tasks on which the model is not fine-tuned. Our findings indicate that Human-in-the-loop Question Decomposition (HQD) can

*Equal Contribution

[1]https://github.com/Pruthvi98/QuestionDecomposition

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4553 - 4569
December 7-11, 2022

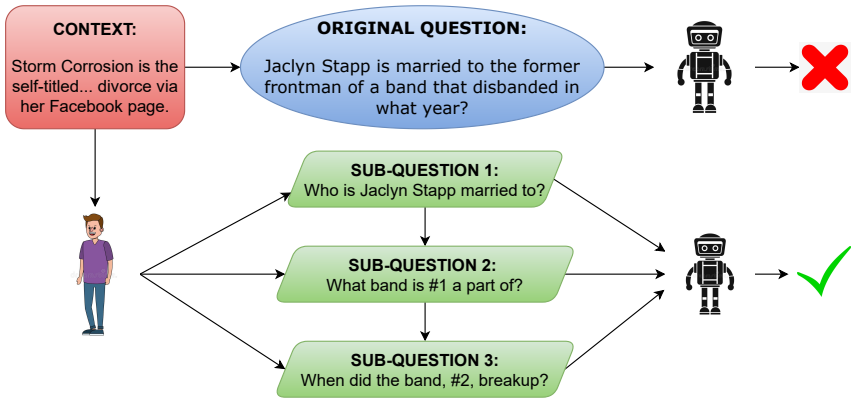

Figure 1: The original question is answered incorrectly by a model. A human then decomposes the question into a set of simpler questions which the model then answers correctly.

potentially provide an alternate path to building large LMs. We hope our work will encourage the community to develop human-centric solutions that actively involve humans while leveraging NLP resources.

2 Related Work

A recent methodology to reason over multiple sentences in reading comprehension datasets is to decompose the question into single-hop questions (Talmor and Berant, 2018; Min et al., 2019). Min et al. (2019) decompose questions from HOTPOTQA using span predictions based on reasoning types and picks the best decomposition using a decomposition scorer. Khot et al. (2020b) generate decompositions by training a BART model on question generation task by providing context, answers and hints. Wolfson et al. (2020) crowd-sourced annotations for decompositions of questions. Perez et al. (2020), on the other hand, uses the unsupervised mechanism of generating decomposition by mapping a hard question to a set of candidate sub-questions from a question corpus. Iyyer et al. (2017) answer a question sequentially using a neural semantic parsing framework over crowdsourced decompositions for questions from WikiTableQuestions. Decomposition using text-to-SQL query conversion has also been studied (Guo et al., 2019). Also, knowledge graphs are combined with neural networks to generate decompositions (Gupta and Lewis, 2018). Recently, Xie et al. (2022) presented another use case where decompositions can be used to probe models to create explanations for their reasoning.

Name	Type
HOTPOTQA	Multihop RC
DROP	Mulithop RC
STRATEGYQA	Strategic Reasoning
MULTIRC	RC
BREAK	RC
MATHQA	Mathematical Reasoning
QASC	Fact-based Multichoice
SVAMP	Context-based Math Word Problems

Table 1: Type of QA task corresponding to each dataset. RC: Reading Comprehension

3 Methods

3.1 Datasets

We select eight datasets covering a diverse set of reasoning skills and domains: (1) HOTPOTQA (Yang et al., 2018), (2) DROP (Dua et al., 2019), (3) MULTIRC (Khashabi et al., 2018), (4) STRATEGYQA (Geva et al., 2021), (5) QASC (Khot et al., 2020a), (6) MATHQA (Amini et al., 2019), (7) SVAMP (Patel et al., 2021), and (8) BREAK (Wolfson et al., 2020). Table 1 indicates the different task types for each dataset.

3.2 Decomposition Process

For each dataset, we randomly select 50 instances for manual decomposition. The question in each dataset is decomposed into two or more questions. Table 2, 3, 4 and 5 show examples of decomposition for various datasets. For each dataset, we created a set $\mathcal{D}$ for decomposed questions. Each element $\mathcal{D}_i \in \mathcal{D}$ can be represented as below:

$$\mathcal{D}_i = \{\mathcal{C}_i, \mathcal{Q}_i, \mathcal{Q}_d, \mathcal{A}_i, \mathcal{A}_d\},$$

where $\mathcal{C}_i$ is the context paragraphs, $\mathcal{Q}_i$ is the original question, $\mathcal{Q}_d$ is the set of decomposed ques-

tions, $\mathcal{A}_i$ is an original answer, and $\mathcal{A}_d$ is the set of answers for corresponding decomposed questions. For questions that require arithmetic or logical operations, we use a computational unit as suggested in Khot et al. (2020b), which takes a decomposed question as input in the following format:

$$\{\mathcal{O}\}!\#m_1!\#m_2!....!\#m_n,$$

where $\mathcal{O}$ = {summation, difference, division, multiplication, greater, lesser, power, concat, return, remainder}, $\#m_i$ are answers of previous decomposed questions and ! separates the operands.

4 Experimental Setup

Models We use GPT-3 (Brown et al., 2020) to generate answers for original and decomposed questions. To show that QD significantly improves performance even on simpler models, we use RoBERTa-base finetuned on SQuAD 2.0 dataset (i.e., RoBERTa-SQuAD). Additionally, we use RoBERTa-base finetuned on BoolQ dataset (Clark et al., 2019) (i.e., RoBERTa-BoolQ) for original and decomposed questions in StrategyQA since they are True/False type questions.

Experiments To create baselines, we evaluate all models on the original question along with the context. We evaluate all models on the manually decomposed questions in the proposed method. We carry out all experiments in GPT-3 by designing prompts for each dataset[2]. For RoBERTa-based models, we use RoBERTa-SQuAD for MultiRC, Break, HotpotQA and DROP datasets, since SQuAD 2.0 is designed for a reading comprehension task. For StrategyQA, we use two RoBERTa-base models: (1) RoBERTa-BoolQ, which is used to answer the final boolean type of questions, and (2) RoBERTa-SQuAD which is used to answer the remaining decomposition questions. For SVAMP, we use the RoBERTa-SQuAD model to extract the necessary operands using decomposed questions and then we use the computational module to perform various operations. In all experiments, we use decomposition to get to the final answer sequentially.

Metrics For all our experiments, we use Rouge-L (Lin, 2004), F_1-score and Exact Match (EM) as the evaluation metrics.

[2]See Appendix A for more details

5 Results and Analysis

Here, we divide our datasets into four categories: (1) RC: HotpotQA, DROP, MultiRC, and Break in Reading Comprehension (RC), (2) MATH: MathQA and SVAMP in Mathematical reasoning , (3) MC: QASC in Multi-Choice QA (MC) , and (4) SR: StrategyQA in Strategy Reasoning (SR). All results presented in this sections are averaged over tasks for each category.

5.1 Experimental Results

GPT-3 Figure 3 shows the GPT-3 performance in terms of average F_1-scores for each category. From the Figure 3, we can observe that our proposed approach outperforms baseline by $\sim 24\%$. Appendix D presents all results in terms of F_1-scores, EM and Rouge-L for all datasets and categories.

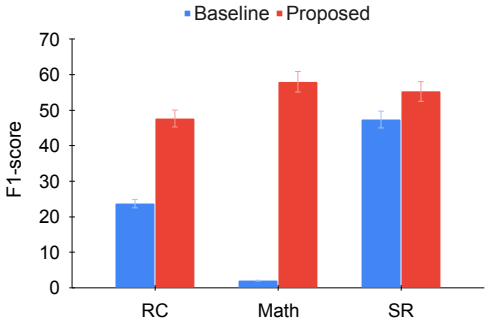

Figure 2: Results in terms of F_1-score across different categories for RoBERTa-based models. RC: Reading Comprehension, MATH: Mathematical reasoning, SR: Strategy Reasoning.

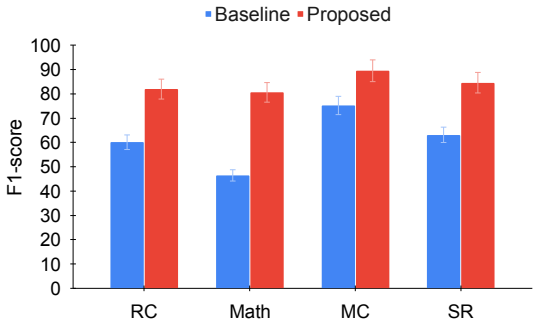

Figure 3: Results in terms of F_1-score across different categories for GPT-3. RC: Reading Comprehension, MATH: Mathematical reasoning, MC: Multi-Choice QA, SR: Strategy Reasoning.

RoBERTa Figure 2 represents the results we obtain using RoBERTa-based models in terms of

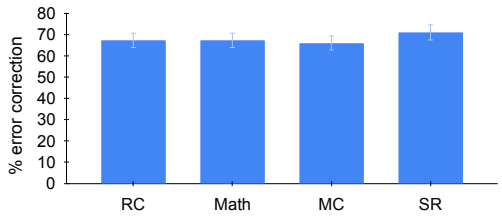

Figure 4: % error correction by using decompositions with GPT3

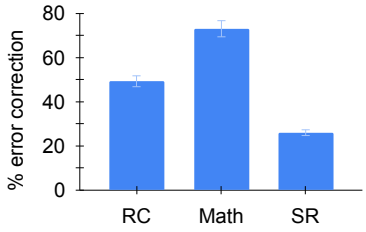

Figure 5: % error correction by using decompositions with RoBERTa

F_1-scores for each category. On an average, we achieve $\sim$ 29% of significant improvement compared to the baseline. Appendix D presents all results in terms of F_1-scores, EM and Rouge-L for all datasets and categories.

5.2 Analysis

Customized Question Decomposition for Each Model There can be multiple ways to decompose a question based on the context. Multiple factors go into deciding how to break down a question. One factor is the strength of the model. For instance, if we use a model finetuned on SQuAD, it might be beneficial to ensure that the decompositions are more granular and are generated to answer from a context span. On the other hand, if we have a more sophisticated model like GPT3, we might not necessarily need to do so. The results shown in Figure 2 are obtained on RoBERTa finetuned on SQuAD by using decompositions originally designed for GPT3; note that in this case, the answers to the decompositions might not always be the span of a particular sentence in the context. However, we achieve a decent performance improvement. We believe the performance gain will be greater if decompositions are designed to match the model's strengths. Examples of such decompositions are included in the Appendix A.

Qualitative Analysis We conduct qualitative analysis to capture the evaluation aspects missed in the automated evaluation metrics. Here, we manually inspect and consider a generated answer to be correct if it is semantically similar to the gold annotation. Figure 4 and 5 show the contribution of QD in correcting model prediction. We observe that the decompositions correct more than 60% of the errors made on the original questions.

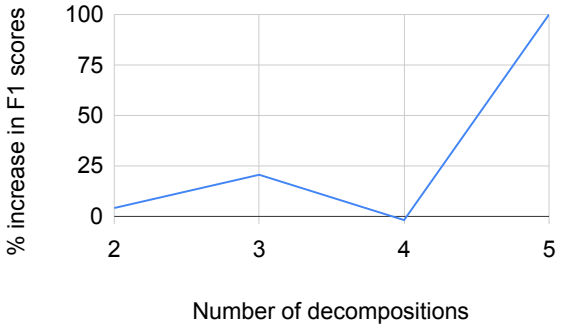

Figure 6: Performance improvements in F1 scores for questions with 2, 3, 4 and 5 decompositions.

Error Analysis We conduct error analysis and observe that the major source of error is the error propagated from one of the decomposed questions. Errors, in general, are of two types: (i) incorrect span selection and (ii) failure to collect all possible answers in the initial step of decomposition; this often omits the actual correct answer leaving no room for later decomposition units to generate the correct answer. Errors occur in QASC because our method of context-independent decomposition (via intuition) sometimes leads to open-ended questions which models find hard to answer. Examples of errors have been included in the Appendix B.

Effect of Decomposition on Math Datasets We observe that Math datasets benefit the most from decomposition. This may be because of two reasons: 1) majority of math questions can be decomposed as a combination of extractive QA (where the answer is a span) and a symbolic calculation. Both of these are strengths of language models (note that we use calculators that provide accurate answers consistently). However, this is not necessarily true in case of other QA tasks. In a decomposition chain, if the answer in one step goes wrong, it propagates till the end and the final prediction becomes wrong. 2) language models by default struggle to do math tasks (Patel et al., 2021; Mishra et al., 2022), so the performance improvement seems more prominent there.

Effect of Number of Decompositions on Results We typically decompose a question based on the number of operations associated with it (e.g. mathematical calculation or single hop operation). Increase in the number of decompositions has the advantage that it simplifies the original question, but it can also have the disadvantage that if the answer to one of the questions in the chain is incorrect, the end answer becomes incorrect. This is also evident from our empirical analysis on HOTPOTQA and SVAMP datasets where we observe that there is no direct correlation between the number of labeling QA and the final performance. Figure 6 shows the variation in model performance improvement observed for questions with 2, 3, 4 and 5 decompositions.

Efforts to Automate Decomposition For HOTPOTQA, DROP, and SVAMP, we attempt to automate the decomposition process using GPT3. A limitation for generating decompositions for HOTPOTQA is that the context length makes it difficult to provide sufficient examples in prompt. With DROP and SVAMP, we observe that GPT-3 often generates incorrect arithmetic operations for the last sub-question. It also often fails to develop coherent decompositions of the questions. We also finetune a BART-base (Lewis et al., 2020) model on our handwritten decompositions. However, the model overfits and fails to produce meaningful decompositions, probably due to the limited number of training samples (see Appendix C for examples, details and results).

6 Conclusion

The recent trend of building large LMs may not be sustainable to solve evolving benchmarks. We believe that modifying data samples can significantly help the model improve performance. We study the effect of Question Decomposition (QD) on a diverse set of tasks. We decompose questions manually and significantly improve model performance (24% for GPT3 and 29% for RoBERTa-SQuAD along with a symbolic calculator). Our findings indicate that Human-in-the-loop Question Decomposition (HQD) can potentially provide an alternate path to building large LMs. Our approach provides a viable option to involve people in NLP research. We hope our work will encourage the community to develop human-centric solutions that actively involve humans while leveraging NLP resources.

Limitations

Our human-in-the-loop methodology shows promising results by decomposing questions, however, certain questions are still difficult to decompose for humans as well. For instance, the question *"Which country is New York in?"*, is hard to decompose further. Determining which questions to decompose is also an important challenge and under-explored in this work. Furthermore, decomposed questions in the chain which have more than one correct answers might lead to an incorrect final answer. Automating the process of decomposition while addressing these issues is a promising area for future work.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *ArXiv*, abs/1905.13319.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Dheeru Dua, Yizhong Wang Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *CoRR*, abs/1903.00161.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and Dongmei Zhang. 2019. Towards complex text-to-sql in cross-domain database with intermediate representation. *arXiv preprint arXiv:1905.08205*.

Nitish Gupta and Mike Lewis. 2018. Neural compositional denotational semantics for question answering. *arXiv preprint arXiv:1808.09942*.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In *NAACL*.

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

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2020b. Text modular networks: Learning to decompose tasks in the language of existing models. *CoRR*, abs/2009.00751.

Kirby Kuznia, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. Less is more: Summary of long instructions is better for program synthesis. *arXiv preprint arXiv:2203.08597*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

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*.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. *arXiv preprint arXiv:1906.02916*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2021a. Reframing instructional prompts to gptk's language. *ACL Findings*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021b. Cross-task generalization via natural language crowdsourcing instructions. *ACL*.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3505–3523.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, M Hassan Murad, and Chitta Baral. 2022. In-BoXBART: Get Instructions into Biomedical Multi-Task Learning. *NAACL 2022 Findings*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. 2020. Unsupervised question decomposition for question answering. *arXiv preprint arXiv:2002.09758*.

Ravsehaj Singh Puri, Swaroop Mishra, Mihir Parmar, and Chitta Baral. 2022. How many data samples is an additional instruction worth? *arXiv preprint arXiv:2203.09161*.

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021a. Multitask prompted training enables zero-shot task generalization.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021b. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*, 8:183–198.

Kaige Xie, Sarah Wiegreffe, and Mark Riedl. 2022. Calibrating trust of multi-hop question answering systems with decompositional probes.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *CoRR*, abs/1809.09600.

A Prompts

Due to the success of large LMs, prompt-based learning is becoming popular to achieve generalization and eliminate the need of creating task-specific models and large scale datasets (Liu et al., 2021). Recently, instructional prompts have been pivotal in improving the performance of LMs and achieving zero-shot generalization (Mishra et al., 2021b; Wei et al., 2021; Sanh et al., 2021b; Wei et al., 2022; Ouyang et al., 2022; Parmar et al., 2022; Puri et al., 2022; Kuznia et al., 2022). We present the instructional prompts that we used to generate answers for various datasets.

A.1 HotpotQA, DROP, Break

Given a context, answer the question using information and facts present in the context. Keep the answer short.
Example:
Input:
Mehmed built a fleet to besiege the city from the sea .Contemporary estimates of the strength of the Ottoman fleet span between about 110 ships , 145 , 160 , 200-250 to 430 . A more realistic modern estimate predicts a fleet strength of 126 ships comprising 6 large galleys, 10 ordinary galleys, 15 smaller galleys, 75 large rowing boats, and 20 horse-transports.:44 Before the siege of Constantinople, it was known that the Ottomans had the ability to cast medium-sized cannons, but the range of some pieces they were able to field far surpassed the defenders' expectations. Instrumental to this Ottoman advancement in arms production was a somewhat mysterious figure by the name of Orban , a Hungarian .:374 One cannon designed by Orban was named "Basilica" and was 27 feet long, and able to hurl a 600lb stone ball over a mile .
Question: How many ordinary galleys and large rowing boats is estimated from the fleet strength?
Output:
Answer: 85

Input:
Context: «CONTEXT»
Question: «QUESTION»
Output:
Answer: «OUTPUT GENERATED BY GPT3»

HotpotQA	***Context***: The Larkspur Press is a small letter-press publisher based in Monterey, Kentucky .... , The film also features appearances by Helen Keller, Anne Sullivan, Kate Adams Keller and Phillips Brooks Keller as themselves. The movie was directed by George Foster Platt and written by Francis Trevelyan Miller. ***Original Question***: Are John O'Hara and Rabindranath Tagore the same nationality? ***True Answer***: no ***Decomposed Question 1***: What is John O'Hara's nationality? ***Generated Answer***: American ***Decomposed Question 2***: What is Rbindranath Tagore's nationality? ***Generated Answer***: Indian ***Decomposed Question 3***: Is #1 and #2 the same nationality? ***Generated Answer***: No
DROP	***Context***: Mehmed built a fleet to besiege the city from the sea .... and able to hurl a 600 lb stone ball over a mile . ***Original Question***: How many ordinary galleys and large rowing boats is estimated from the fleet strength? ***True Answer***: 85 ***Decomposed Question 1***: How many ordinary galleys were there? ***Generated Answer***: 10 ***Decomposed Question 2***: How many large rowing boats were there? ***Generated Answer***: 75 ***Decomposed Question 3***: summation ! #1 ! #2 ***Generated Answer***: 85

Table 2: Examples for DROP and HotpotQA.

A.2 MATHQA

Prompt for the original question:
Given a problem and 5 options, return the correct option. In order to choose the correct option, you will have to perform some mathematical operations based on the information present in the problem. Look at the examples given below to understand how to answer.

Input: Problem: the volume of water inside a swimming pool doubles every hour . if the pool is filled to its full capacity within 8 hours , in how many hours was it filled to one quarter of its capacity
Options: a) 2, b) 4, c) 5, d) 6, e) 7
Output:
Answer: 6

Input:
Problem: a train 200 m long can cross an electric pole in 5 sec and then find the speed of the train ?
Options: a) 114 , b) 124 , c) 134 , d) 144 , e) 154
Output:
Answer: 144

Input:
Problem: «Problem»
Options: «options»
Output:
Answer: «OUTPUT GENERATED BY GPT3»

A.3 SVAMP

Prompt used for both decomposed questions and original questions. The examples contain both decomposed type questions and original type questions.

Given some context, answer a given question. Use the examples given below as reference.

Example 1:
Input:
Context: It takes 4.0 apples to make 1.0 pie.
Question: How many apples does it take to make 504.0 pies?
Output:
Answer: 2016

Example 2:
Input:
Context: Mary is baking a cake.The recipe calls for 7.0 cups of flour and 3.0 cups of sugar.She already put in 2.0 cups of flour.

Question: How many cups of flour did recipe called?
Output:
Answer: 7

Example 3:
Input:
Context: Each pack of dvds costs 76 dollars. If there is a discount of 25 dollars on each pack
Question: How much is each pack of dvds without the discount?
Output:
Answer: 76

Example 4:
Input:
Context: Conner has 25000.0 dollars in his bank account.Every month he spends 1500.0 dollars.He does not add money to the account.
Question: How many dollars Conner spends every month?
Output:
Answer: 1500

Input:
Context: <CONTEXT»
Question: «QUESTION»
Output:
Answer: «OUTPUT GENERATED BY GPT3»

A.4 StrategyQA

Input:
Context: A melodrama is a dramatic work ...The passengers' response to the hijacking has come to be invested with great moral significance.
Question: What do tearjerkers refer to?
Output:
Answer: a story, song, play, film, or broadcast that moves or is intended to move its audience to tears.

Input:
Context: The purpose of the course is learning to soldier as ... The main motor symptoms are collectively called "parkinsonism", or a "parkinsonian syndrome".
Question: True or False: Could someone experiencing A tremor, or shaking, Slowed movement (bradykinesia), Rigid muscles, Impaired posture and balance, Loss of automatic movements, Speech changes, Writing changes. complete Volunteer for assignment and be on active duty. Have a General Technical (GT) Score of 105 or higher
Output:
Answer: False

Input:
Context: The Scientific Revolution was a series of events that marked the emergence of modern science during the early modern period, ...The first-generation iPhone was released on June 29, 2007, and multiple new hardware iterations with new iOS releases have been released since.
Question: True or False: Did 1543 occur before 2007?
Output:
Answer: False

Input:
Context: «CONTEXT»
Question: «QUESTION»
Output:
Answer: «OUTPUT GENERATED BY GPT3»

A.5 QASC

Prompt for original question:

Answer the given question. The question contains options A-H, choose and return the correct option. Look at the examples given below.

Input:
What are the vibrations in the ear called? (A) intensity (B) very complex (C) melanin content (D) lamphreys (E) Otoacoustic (F) weater (G) Seisometers (H) trucks and cars
Output:
Answer: Otoacoustic

Input:
«QUESTION»
Output:
Answer: «OUTPUT GENERATED BY GPT3»

Prompt for decomposed question:

Given a yes or no question, return yes if the answer is yes. Otherwise return no.
«QUESTION»
Answer: «OUTPUT GENERATED BY GPT3»

SVAMP	***Context***: Bryan took a look at his books as well.If Bryan has 56.0 books in each of his 9.0 bookshelves. ***Original Question***: How many books does he have in total? ***Answer***: 504.0 ***Decomposed Question 1***:How many books in each bookshelf? ***Answer***: 56.0 ***Decomposed Question 2***:How many bookshelves? ***Answer***: 9.0 ***Decomposed Question 3***: multiplication ! #1 ! #2 ***Answer***: 504.0
MATHQA	***Problem***: if a train , travelling at a speed of 180 kmph , crosses a pole in 6 sec , then the length of train is ? ***Options***: a) 300 , b) 125 , c) 288 , d) 266 , e) 121 ***Annotated Formula***: multiply(multiply(180, const_0.2778), 6) ***Answer***: 300 ***Generated Answer***: 266 ***Decomposed Question 1***: multiplication ! 0.2778 ! 180 ***Answer***: 50.004 ***Decomposed Question 2***: multiplication ! 50.004 ! 6 ***Answer***: 300

Table 3: Decomposition Examples for SVAMP and MathQA. We use the annotated formula presented in the dataset to make our decompositions.

StrategyQA	***Context***: Mail carriers, also referred to as mailmen or letter carriers, ... Clothing also provides protection from ultraviolet radiation. ***Original Question***: True or False: Mail carriers need multiple uniforms. ***Original Answer***: True ***Generated Answer***: False ***Decomposed Question 1***: What seasons do mail carriers work through? ***Generated Answer***: All seasons ***Decomposed Question 2***: True or False: In order to make it through all of #1, one needs multiple clothing pieces. ***Generated Answer***: True
QASC	***Original Question***: what kind of beads are formed from vapor condensing? (A) h2o (B) H20 (C) tiny (D) carbon (E) hydrogen (F) rain (G) oxygen (H) Dew ***Answer***: h2o ***Decomposed Question 1***: Are #1 beads formed from vapor condensing? ***Answer***: yes

Table 4: Examples of decompositions for StrategyQA and QASC datasets. For each option in QASC, #1 is replaced with the option and posed to GPT-3 as a yes or no question.

A.6 MultiRC

Given a context-question pair, answer the question using information and facts present in the context. Keep your answers as short as possible.

Example:
Input:
Context: Should places at the same distance from the equator have the same climate? You might think they should. Unfortunately, you would not be correct to think this. Climate types vary due to other factors besides distance from the equator. So what are these factors? How can they have such a large impact on local climates? For one thing, these factors are big. You may wonder, are they as big as a car. Think bigger. Are they bigger than a house? Think bigger. Are they bigger than a football stadium? You are still not close. We are talking about mountains and oceans. They are big features and big factors. Oceans and mountains play a huge role in climates around the world. You can see this in Figure above. Only one of those factors is latitude, or distance from the equator.
Question: Name at least one factor of climate
Output:
Answer: Oceans

Example:
Input:
Context: Earth processes have not changed over time. The way things happen now is the same way things happened in the past. Mountains grow and mountains slowly wear away. The same process is at work the same as it was billions of years ago. As the environment changes, living creatures adapt. They change over time. Some organisms may not be able to adapt. They become extinct. Becoming extinct means they die out completely. Some geologists study the history of the Earth. They want to learn about Earths past. They use clues from rocks and fossils. They use these clues to make sense of events. The goal is to place things in the order they happened. They also want to know how long it took for those events to happen.
Question: What is one example of how the earth's processes are the same today as in the past?
Output:
Answer: Things develop and then wither away

Input:
Context:: «CONTEXT»
Question: «QUESTION»
Output:
Answer: <ANSWER GENERATED BY GPT3»

B Error Examples

This section discusses the errors generated by using decompositions. We observe two types of errors while answering decomposed questions. The final answer is wrong because previous sub-questions were answered incorrectly either because such a question has multiple correct answer, or simply because the model could not understand the question correctly.

Context: ... Roger David Casement (1 September 1864 - 3 August 1916), formerly known as Sir Roger Casement .... In collaboration with Roger Casement, Morel led a campaign against slavery in the Congo Free State, founding the Congo Reform Association .... The association was founded in March, 1904, by Dr. Henry Grattan Guinness (1861-1915), Edmund Dene Morel, and Roger Casement ...
Question:
When was the date of birth of one of the founder of Congo Reform Association?
True Answer: 1 September 1864
Generated Answer: 18 October 1914

Decomposed Question 1:
Who is the founder of the Congo Reform Association?
True Answer: Roger Casement
Generated Answer: Henry Grattan Guinness

Decomposed Question 2: When was #1 born?
True Answer: 1 September 1864
Generated Answer: 1861

Above is an example from HotpotQA. As can be seen from the context, Congo Reform Association had multiple founders. GPT3 did give a correct answer among a set of correct answers whereas the ground truth answer provided by the dataset was some other correct option.

Below is an example of incorrect retrieval. The answer generated for the first decomposed question incorrectly returns cities taken by Ottomans as well instead of just the Venetians. Hence, the final decomposed questions returns the incorrect count.

MultiRC	***Context***: Sometimes a full Moon moves through Earths shadow. ... The Moon glows with a dull red coloring during a total lunar eclipse. ***Original Question***: Is it more common for the Moon to travel completely in the Earth's umbra or only partially? ***List of correct answers***: Partially, A total eclipse is less common than partial so it is more common for the moon to travel partially in Earth's umbra ***Decomposed Question 1***:When does the Moon travel's completely in Earth's umbra? ***Answer***: total lunar eclipse ***Decomposed Question 2***:When does the Moon travel's partially in Earth's umbra? ***Answer***: partial lunar eclipse ***Decomposed Question 3***: Which is more common #1 or #2? ***Answer***: partial lunar eclipse ***Decomposed Question 4***: Does the Moon travel partially or completely in #3? ***Answer***: partially

Table 5: Decomposition Examples for MultiRC. MultiRC has multiple correct answer and the final correct answer which gives the best metrics for the generated answer is chosen as the correct answer corresponding to the generated answer.

Context: In the Morean War, the Republic of Venice besieged Sinj in October 1684 and then again March and April 1685, but both times without success. In the 1685 attempt, the Venetian armies were aided by the local militia of the Republic of Poljica, who thereby rebelled against their nominal Ottoman suzerainty that had existed since 1513. In an effort to retaliate to Poljica, in June 1685, the Ottomans attacked Zadvarje, and in July 1686 Dolac and Srijane, but were pushed back, and suffered major casualties. With the help of the local population of Poljica as well as the Morlachs, the fortress of Sinj finally fell to the Venetian army on 30 September 1686. On 1 September 1687 the siege of Herceg Novi started, and ended with a Venetian victory on 30 September. Knin was taken after a twelve-day siege on 11 September 1688. The capture of the Knin Fortress marked the end of the successful Venetian campaign to expand their territory in inland Dalmatia, and it also determined much of the final border between Dalmatia and Bosnia and Herzegovina that stands today. The Ottomans would besiege Sinj again in the Second Morean War, but would be repelled. On 26 November 1690, Venice took Vrgorac, which opened the route towards Imotski and Mostar. In 1694 they managed to take areas north of the Republic of Ragusa, namely Čitluk, Gabela, Zažablje, Trebinje, Popovo, Klobuk and Metković. In the final peace treaty, Venice did relinquish the areas of Popovo polje as well as Klek and Sutorina, to maintain the pre-existing demarcation near Ragusa.
Question:
How many cities did Venice try to take?
True Answer: 10
Generated Answer: 3

Decomposed Question 1:
Which cities did Venice try to take?
True Answer: Sinj, Knin, Vrgorac, Čitluk, Gabela, Zažablje, Trebinje, Popovo, Klobuk and Metković
Generated Answer: Sinj, Zadvarje, Dolac, Srijane, Knin, Vrgorac, Čitluk, Gabela, Zažablje, Trebinje, Popovo, Klobuk and Metković

Decomposed Question 2: What is the count of the cities mentioned in #1?
True Answer: 10
Generated Answer: 14

The samples for QASC are provided without context. Without the context, the answers to some of the decomposed questions can be open ended. Certain options can be unambiguously wrong and some are unambiguously correct. Below is an example:
Question: What can knowledge of the stars be used for? (A) travel (B) art (C) as a base (D) safety (E) story telling (F) light source (G) vision (H) life
True Answer: travel
Generated Answer: art

Decomposed Question: Can the knowledge

of stars be used for the following: #?

The decomposed question for each option is posed as a yes or no question to GPT3. It returns yes for art and story telling but not for travel.

C Examples, Results and Details for Automation

We attempt to automate the process of decomposition using GPT3. We use the examples from manual decomposition in the prompts given to GPT3, some of which are presented below. The results obtained from the experiments are presented in Table 6. The generated decompositions are answered using RoBERTa-base finetuned on SQuAD 2.0 dataset.

In this section, we present the prompts we used while attempting to automatically generate decomposed questions using GPT3.

The prompt for generating decompositions for DROP was as follows:

Decompose a given question by breaking it into simpler sub-questions. The answer to each subsequent sub-question should lead towards the answer of the given question. To do so, use the context provided and look at the examples. Here are some helpful instructions:

1. If the given question compares two things, best strategy is to generate sub-questions that finds the answer to each of those things and compare them in the last sub-question.

2. Some sub-questions must contain phrases like "answer of sub-question 1".

3. If a sub-question is an arithmetic operation, then the sub-question should be framed as operation ! "answer of sub-question 1" ! "answer of sub-question 2".

4. The operation used in 3) is always one of the following: summation, difference, greater, lesser.

Example 1:
Context: Mehmed built a fleet to besiege the city from the sea .Contemporary estimates of the strength of the Ottoman fleet span between about 110 ships , 145 , 160 , 200-250 to 430 . A more realistic modern estimate predicts a fleet strength of 126 ships comprising 6 large galleys, 10 ordinary galleys, 15 smaller galleys, 75 large rowing boats, and 20 horse-transports.:44 Before the siege of Constantinople, it was known that the Ottomans had the ability to cast medium-sized cannons, but the range of some pieces they were able to field far surpassed the defenders' expectations. Instrumental to this Ottoman advancement in arms production was a somewhat mysterious figure by the name of Orban , a Hungarian. One cannon designed by Orban was named Bäsilicaänd was 27 feet long, and able to hurl a 600 lb stone ball over a mile .
Question: How many ordinary galleys and large rowing boats is estimated from the fleet strength?
Sub-question 1: How many ordinary galleys were there?
Sub-question 2: How many large rowing boats were there?"
Sub-question 3: summation ! "answer of sub-question 1" ! "answer of sub-question 2"

Example 2:
Context: As of the census of 2000, there were 14,702 people, 5,771 households, and 4,097 families residing in the county. The population density was 29 people per square mile (11/km²). There were 7,374 housing units at an average density of 14 per square mile (6/km²). The racial makeup of the county was 98.02% Race (United States Census), 0.69% Race (United States Census) or Race (United States Census), 0.35% Race (United States Census), 0.11% Race (United States Census), 0.05% Race (United States Census), 0.08% from Race (United States Census), and 0.71% from two or more races. 0.44% of the population were Race (United States Census) or Race (United States Census) of any race.
Question: How many more people than households are reported according to the census?
Sub-question 1: As of the 2000 census, how many people are residing in the country?
Sub-question 2: As of the 2000 census, how many households are reported?
Sub-question 3: difference !"answer of sub-question 1" ! "answer of sub-question 2"

Example 3: Context: As of the census of 2000, there were 49,129 people, 18,878 households, and 13,629 families residing in the county.

Dataset	F1		EM		Rouge-L	
	Baseline	Decompose	Baseline	Decompose	Baseline	Decompose
HotpotQA	**32.68**	14.12	**29.50**	11.47	**33.29**	14.00
DROP	**22.8**	3.77	**21.69**	3.77	**23.4**	3.76
SVAMP	7.4	**17.35**	7.4	**17.35**	7.4	**17.35**
Average	**20.96**	11.74	**19.53**	10.86	**21.36**	11.70

Table 6: Results obtained by using decomposed questions generated using GPT3

The population density was 88 people per square mile (34/km2). There were 21,779 housing units at an average density of 39 per square mile (15/km2). The racial makeup of the county was 74.4% Race (United States Census), 20.4% Race (United States Census) or Race (United States Census), 0.60% Race (United States Census), 1.1% Race (United States Census), 0.15% Race (United States Census), 1.3% from Race (United States Census), and 2.2% from two or more races. 3.4% of the population were Race (United States Census) or Race (United States Census) of any race. 2.85% of the population reported speaking Spanish language at home, while 1.51% speak German language. Question: How many more people are there than families? Sub-question 1: How many people are there in the 2000 census? Sub-question 2: How many families are recorded in the 200 census? Sub-question 3: difference ! "answer of sub-question 1" ! "answer of sub-question 2"

Context: «CONTEXT»
Question: «QUESTION»

«OUTPUT GENERATED BY GPT3»

The prompt for HotpotQA was similar, except replacing the examples with instances from HotpotQA. For SVAMP, since the context was much smaller, we could give more examples. The prompt for SVAMP is as shown below:

Decompose a given question by breaking it into simpler sub-questions. The answer to each subsequent sub-question should lead towards the answer of the given question. To do so, use the context provided and look at the examples.

Here are some helpful instructions:

1. If the given question compares two things, best strategy is to generate sub-questions that finds the answer to each of those things and compare them in the last sub-question,
   2) Some sub-questions must contain phrases like "answer of sub-question 1".

2. Some sub-questions must contain phrases like "answer of sub-question 1".

3. If a sub-question is an arithmetic operation, then the sub-question should be framed as operation ! "answer of sub-question 1" ! "answer of sub-question 2".

4. The operation used in 3) is always one of the following: summation, difference, greater, lesser

Example 1:
Context: Jessica had 8.0 quarters in her bank . Her sister borrowed 3.0 of her quarters. How many quarters does Jessica have now?
Sub-question 1: How many quarters did Jessica have in her bank initially?
Sub-question 2: How many quarters did Jessica's sister borrow?
Sub-question 3: difference ! "answer of sub-question 1" ! "answer of sub-question 2"

Example 2:
Context: Shawn has 13.0 blocks. Mildred has with 2.0 blocks. Mildred finds another 84.0. How many blocks does Mildred end with?
Sub-question 1: How many blocks does Mildred start with?
Sub-question 2: How many blocks does Mildred find?
Sub-question 3: summation ! "answer of sub-question 1" ! "answer of sub-question 2"

Example 3:
Context: Dave was helping the cafeteria workers

pick up lunch trays, but he could only carry 9.0 trays at a time. If he had to pick up 17.0 trays from one table and 55.0 trays from another. how many trips will he make?
Sub-question 1: How many trays did Dave have to pick up from the first table?
Sub-question 2: How many trays did Dave have to pick up from the second table?
Sub-question 3: summation ! "answer of sub-question 1" ! "answer of sub-question 2"
Sub-question 4: How many lunch trays could Dave carry at a time?
Sub-question 5: division ! "answer of sub-question 3" ! "answer of sub-question 4"

Example 4:
Context: Paco had 93.0 cookies. Paco ate 15.0 of them. How many cookies did Paco have left?
Sub-question 1: How many cookies did Paco start with?
Sub-question 2: How many cookies did Paco eat?
Sub-question 3: difference ! "answer of sub-question 1" ! "answer of sub-question 2"

Example 5:
Context: 43 children were riding on the bus. At the bus stop some children got off the bus. Then there were 21 children left on the bus. How many children got off the bus at the bus stop?
Sub-question 1: How many children were on the bus at the beginning?
Sub-question 2: How many children were left on the bus?
Sub-question 3: difference ! "answer of sub-question 1" ! "answer of sub-question 2"

Example 6:
Context: 28 children were riding on the bus. At the bus stop 82 children got on the bus while some got off the bus. Then there were 30 children altogether on the bus. How many more children got on the bus than those that got off?
Sub-question 1: How many children were on the bus at the beginning?
Sub-question 2: How many children were left on the bus?
Sub-question 3: difference ! "answer of sub-question 1" ! "answer of sub-question 2"

Example 7:
Context: They decided to hold the party in their backyard. If they have 11 sets of tables and each set has 13 chairs, how many chairs do they have in the backyard?
Sub-question 1: How many tables are there in the backyard?
Sub-question 2: How many chairs are on each table?
Sub-question 3: multiplication ! "answer of sub-question 1" ! "answer of sub-question 2"

Context: «CONTEXT + QUESTION»

The examples of decompositions generated for HotpotQA, DROP and SVAMP are shown in Table 7

D Results

We tabulate the results we get for all the datasets for baseline and our proposed mechanism.

DROP	***Context***: Hoping to rebound from their loss to the Patriots, the Raiders stayed at home for a Week 16 duel with the Houston Texans. ... The Texans tried to rally in the fourth quarter as Brown nailed a 40-yard field goal, yet the Raiders' defense would shut down any possible attempt. ***Original Question***: How many yards longer was the longest passing touchdown than the shortest? ***Decomposed Question 1***: What was the length of the shortest touchdown pass? ***Decomposed Question 2***:What was the length of the longest touchdown pass? ***Decomposed Question 3***: greater ! #1 ! #2
DROP	***Context***: In 1085, Guadalajara was retaken by the Christian forces of Alfonso VI . The chronicles say that the Christian army was led by Alvar Fanez de Minaya, one of the lieutenants of El Cid. From 1085 until the Battle of Las Navas de Tolosa in 1212, the city suffered wars against the Almoravid and the Almohad Empires. In spite of the wars, the Christian population could definitely settle down in the area thanks to the repopulation with people from the North who received their first fuero in 1133 from Alfonso VII.In 1219, the king Fernando III gave a new fuero to the city .During the reign of Alfonso X of Castile, the protection of the king allowed the city to develop its economy by protecting merchants and allowing markets. ***Original Question***: When did the first battle against Guadalajara take place? ***Decomposed Question 1***: When was Guadalajara retaken by the Christian forces? ***Decomposed Question 2***:Who led the Christian army? ***Decomposed Question 3***: #1 ! #2

Table 7: Decompositions for DROP generated using GPT3

Dataset	F1		EM		Rouge-L	
	Baseline	Decompose	Baseline	Decompose	Baseline	Decompose
HotpotQA	71.97	**78.53**	70	**76**	73.33	**79.93**
DROP	52.97	**78.16**	46.87	**75.86**	46.72	**77.66**
MultiRC	64.39	**80.74**	33.33	**55.55**	61.24	**77.31**
BREAK	66.81	**84.54**	58	**74**	62.30	**78.56**
Average	60.10	**81.97**	52.64	**76.26**	59.35	**81.10**

Table 8: Comparison of metrics for reading comprehension datasets between GPT3 baseline and Decompose_GPT3

Dataset	F1		EM		Rouge-L	
	Baseline	Decompose	Baseline	Decompose	Baseline	Decompose
MATH	31.1	**82.5**	27.44	**82.22**	23.4	**80.85**
SVAMP	61.80	**78.75**	58.88	**77.5**	55	**77.5**
Average	46.45	**80.62**	43.16	**79.86**	39.2	**79.17**

Table 9: Comparison of metrics for mathematical reasoning datasets between GPT3 baseline and Decompose_GPT3

Dataset	F1		EM		Rouge-L	
	Baseline	Decompose	Baseline	Decompose	Baseline	Decompose
StrategyQA	63.15	**84.61**	63.15	**84.61**	63.15	**84.61**
QASC	75.23	**89.52**	75.23	**89.52**	71.4	**85.71**
Average	69.19	**87.06**	69.19	**87.06**	67.27	**85.16**

Table 10: Comparison of metrics for StrategyQA (strategic reasoning) and QASC (fact-based multichoice) between GPT3 baseline and Decompose_GPT3

Dataset	F1		EM		Rouge-L	
	Baseline	Decompose	Baseline	Decompose	Baseline	Decompose
HotpotQA	32.14	**49.50**	26	**42**	33.33	**50.72**
DROP	25.56	**66.14**	25	**62.5**	25.56	**66.14**
MultiRC	45.74	**48.1**	24.44	**28.88**	44.83	**46.95**
BREAK	24.6	**36.17**	18	**28**	24.31	**35.5**
Average	23.68	**47.65**	20.26	**43.96**	28.76	**50.74**

Table 11: Comparison of metrics for reading comprehension datasets between baseline and decompose settings using RoBERTa-base finetuned on SQuAD.

Dataset	F1		EM		Rouge-L	
	Baseline	Decompose	Baseline	Decompose	Baseline	Decompose
StrategyQA	47.36	**55.26**	47.36	**55.26**	47.36	**55.26**
SVAMP	2	**58**	2	**58**	2	**58**
Average	24.68	**56.63**	24.68	**56.63**	24.68	**56.63**

Table 12: Comparison of metrics for StrategyQA and SVAMP between baseline and decompose settings using RoBERTa-base finetuned on SQuAD. For StrategyQA, RoBERTa-base SQuAD is used to answer intermediate decompositions whereas RoBERTa-base finetuned on BoolQ is used to answer the original question and the final decomposed question