Exploring Task Decomposition for Assisting Large Language Models in Counter-argument Logical Structure Analysis

Wenzhi Wang[1,2] Shoichi Naito[1,2,3] Paul Reisert[4] Naoya Inoue[5,2] Camélia Guerraoui[1,2,6]
Jungmin Choi[2] Irfan Robbani[5] Kentaro Inui[7,1,2]
[1]Tohoku University [2]RIKEN [3]Ricoh Company, Ltd. [4]Beyond Reason
[5]JAIST [6]INSA Lyon [7]MBZUAI
{wang.wenzhi.r7, naito.shoichi.t1, guerraoui.camelia.kenza.q4}@dc.tohoku.ac.jp, beyond.reason.sp@gmail.com
naoya-i@jaist.ac.jp, jungmin.choi@riken.jp, robbaniirfan@jaist.ac.jp, kentaro.inui@mbzuai.ac.ae

Abstract

Counter-**A**rgument **L**ogical **S**tructure **A**nalysis (**CALSA**) is an intricate task that focuses on the automatic analysis of logic patterns of a counter-argument in relation to an initial-argument. It holds significant educational value as informative feedback can be provided based on the analyzed logic pattern. Nevertheless, the complexity of the reasoning skills required for logical structure analysis makes CALSA particularly challenging for current LLMs. To overcome this issue, we explore decomposing the task into several manageable sub-tasks with a pre-defined decision tree and utilize an LLM to reason through the tree. Our experimental results highlight a remarkable improvement in our approach over the baseline, emphasizing the substantial efficacy of our proposed method.

1 Introduction

Counter-arguments (CAs) are a good means to improve the critical-thinking skills of learners, especially given that one has to thoroughly consider the logic of initial arguments (IA) and compose CAs based upon that. In order to maximize learning efficiency, tailored feedback from teachers is extremely valuable, however, it is difficult to provide every learner tailored feedback due to limited human resources and heavy workloads [1]. Therefore, developing a system that can automatically provide constructive feedback to learners' CAs for improving their critical-thinking skills would be a beneficial way of applying artificial intelligence (AI) technology to the educational field.

Toward providing automatic constructive feedback, we consider a two-phase approach in which i) we develop a system that can automatically analyze the logical structure of learners' CAs first, and ii) provide learners with tailored feedback based on the analyzed logical structure. In this work, we address the first phase in the context of debates, based on the logic pattern templates designed by Naito et al. [2], leaving the second phase for future work.

As shown in Figure 1(b), Naito et al. [2] propose the **CA L**ogical **S**tructure **A**nalysis (**CALSA**) task, wherein they create 10 templates to structure the most prevalent logic patterns in CAs in relation to initial arguments that follow a specific argumentation scheme *"Argument from Consequences"* [3]. They construct a dataset consisting of CA essays annotated with the 10 proposed logic pattern templates along with the corresponding slot-fillers. We focus on the CALSA task since the informative template set enables the provision of detailed feedback based on each template, aligning seamlessly with our ultimate goal.

Recent advancements in large language models (LLMs) have facilitated significant progress in the field of computational argumentation, specifically in the case of analyzing the rhetorical relation within an argumentative essay as shown in Figure 1(a) [4]. Nonetheless, it remains a challenge for current LLMs to analyze the underlying logical structure of an argumentative essay in relation to another, as it heavily demands complex reasoning skills which have been reported by various research as one of the weak points of current LLMs [5, 6, 7, 8]. Therefore, we hope to aid LLMs in tackling the intricate CALSA task by decomposing the overall task into several more manageable sub-tasks and utilizing LLMs to address each sub-task without rely-

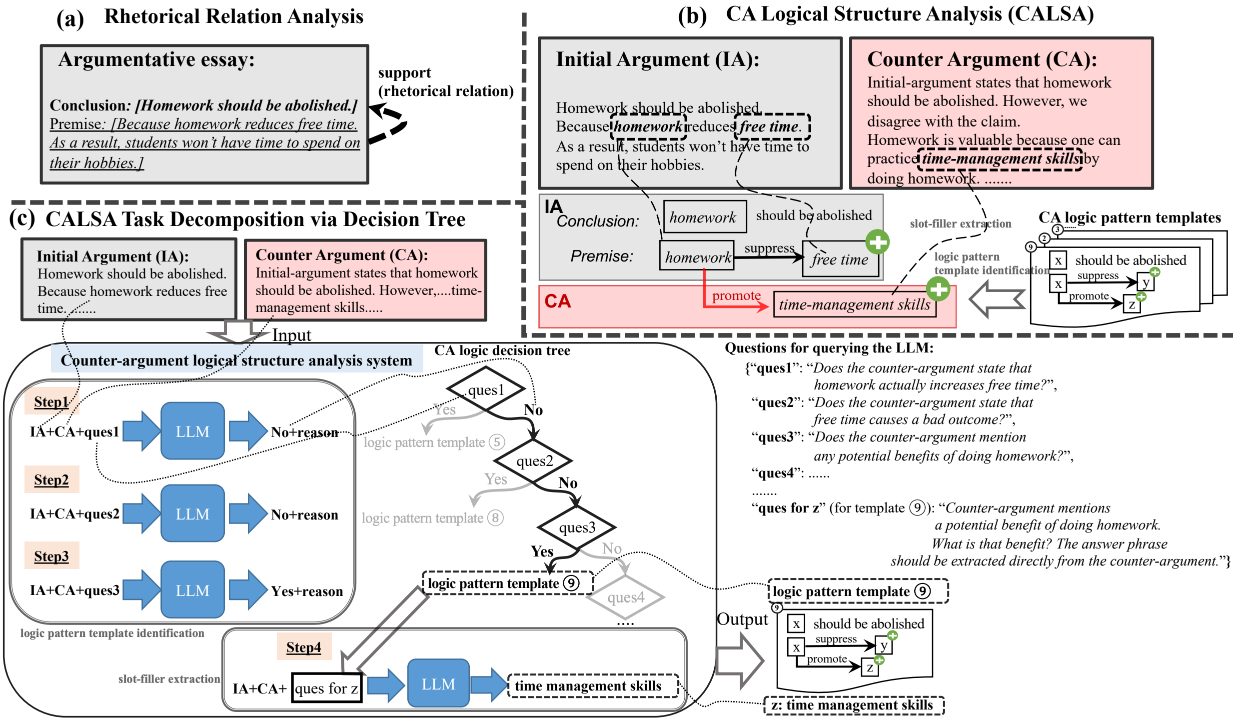


Figure 1 **(a)** The task of analyzing rhetorical relation within an argumentative essay, which includes the identification of argumentative components(i.e. premise, conclusion, etc.) and their relations(support or attack); **(b)** The CALSA task which includes selecting a logic pattern template from the pre-defined templates set and extracting the corresponding slot-fillers from the CA essay. The example illustrates that the CA rebuts IA by introducing another benefit associated with homework. However, it does not explicitly address IA's argument that homework diminishes "free time" which is perceived as a positive thing by the IA; **(c)** An overview of our logical structure analysis system. Given an initial argument (IA) and a counter-argument (CA), our system analyzes the logic pattern of the CA by querying an LLM with the question on a non-leaf node from the pre-defined decision tree at each step. The LLM's responses at each step guide the process along one of the paths in the tree, ultimately leading to a leaf node that represents a logic pattern template. Once the logic pattern template is identified, the LLM is queried again to determine the corresponding slot-filler for the template.

ing on them to solve the entire task in a single step.

Drawing inspiration from recent studies in NLP that explore tree-based approaches for problem-solving [9, 10], as shown in Figure 1(c), we represent the whole process of CA logical structure analysis as a decision tree in which each non-leaf node represents a binary identifiable question that distinguishes a group of logic pattern templates from others, whereas each leaf node represents one of the target logic pattern templates. We utilize an LLM to answer the question on each non-leaf node, by doing so, the LLM will navigate us to a leaf node that represents the final predicted CA logic pattern template. Subsequently, we query the corresponding slot-filler for the identified pattern.

Our proposed approach presents a multitude of benefits, including the division of intricate reasoning tasks into manageable sub-tasks, the interpretability of intermediate reasoning steps, and notably, the control and predictability of the system's output. The latter attribute holds particular significance within an educational context, as it is necessary to govern the content of feedback provided to students based on the CA logical structure analysis results given by the system.

Our experimental results show a notable boost in LLM performance compared to the baseline, which emphasizes the effectiveness of our task decomposition approach.

2 Related work

2.1 LLM's reasoning abilities

Reasoning plays a critical role in human intellectual activities. However, the ability to reason has often been identified as a weak point of language models and other NLP models [5, 6, 7, 8]. Several studies show that NLP models struggle with multiple-step reasoning for in-context learning [5, 7, 6, 11, 12]. With the current advancements of large language models (LLMs), recent research has found that when scaling beyond a certain magnitude of parameters, LLMs start to exhibit exceptional performance on specific reasoning tasks [13]. While these models demonstrate high proficiency in specific reasoning tasks, questions persist regarding whether LLMs are actually reasoning and the extent of their reasoning capabilities [8, 6, 13].

2.2 Problem decomposition

In the field of NLP, several works have exploited the idea of problem decomposition. Some researchers focus on Question-Answering (QA) tasks in which they devise algorithms to automatically break down a challenging question into simpler sub-questions.[14, 15, 16]. Nevertheless, the questions they emphasize are mostly related to factual information, which are inherently easier to automatically decompose. In contrast, the decomposition of our task places a significant emphasis on logical reasoning, making their simplistic automatic approach unsuitable for our context. Another line of work focuses on decomposing the whole task into sub-tasks and prompting LLMs to solve each subtask in order to reach the final answer [17, 18, 19, 9, 10]. Their work, despite the high resemblance to our approach, remains unsuitable for our task due to the lack of control over the final generated results, since they solely rely on prompting LLMs to automatically generate sub-questions without control of the reasoning flow. In this work, we mitigate such an issue by using a pre-defined decision-based parsing tree, as it gives us control over the flow of reasoning steps and the final output of the system.

3 Task Decomposition with Decision Tree

As mentioned previously, we represent the procedure of identifying logic patterns as a decision tree, in which each non-leaf node represents an identifiable binary question that distinguishes a group of logic patterns with the same characteristics from others, and each leaf node represents one of the CA logic pattern templates. Given that CA logic is based on the IA, for each IA, we design a distinct set of questions tailored to the characteristics of each template to query LLMs. Due to space limitations, we show the structure of the decision tree, the questions for each node, and the questions for querying slot-fillers for one IA in Figure 3 in the Appendix.

4 Data

The original dataset proposed by Naito et al. [2] contains 8 unique IAs for 3 different topics and 778 corresponding CAs, each of which has 1 or multiple logic patterns annotated on top. As the Inter-Annotator Agreement for 3 annotators reported for the dataset is moderate, we opt to only use CAs that have annotations agreed by more than 2 annotators. In summary, We utilize one topic (*Should homework be abolished*) which includes 3 different IAs and 160 corresponding unique CAs for our experiments. The statistics are shown in Table 1.

Table 1 The number of CAs utilized in the experiments.

IA ID	#CAs(test)	Main point
HW1	54	HW reduces free time
HW2	53	HW promotes being passive in character
HW4	53	HW promotes incorrect ways of studying
Total	160	

Table 2 Zero-shot precision (P) for CAs in relation to different IAs independently and all combined.

	Baseline		*Decomp*	
IA ID	**P(ptn)**	**P(slots)**	**P(ptn)**	**P(slots)**
HW1	31.5	27.8	**55.6**	**48.1**
HW2	39.6	37.7	**60.4**	**58.5**
HW4	11.3	9.4	**45.3**	**35.8**
ALL	27.5	25	**53.8**	**47.5**

5 Experiments

To test the efficiency of our method, we conduct zero-shot experiments in the following two settings: i) ***Decomp***: we instruct an LLM to address the question on each node step by step. In this scenario, the LLM generates the answer to a question on a non-leaf node (and a question regarding slot-fillers) as well as its explanation of the answer, given an IA, a corresponding CA, and the question. ii) ***Baseline***: we instruct another LLM to solve the entire task in a single step. The LLM is prompted to generate the identifier of the most obvious logic pattern (the mapping between identifiers and the actual logic pattern templates is shown in the prompt) of the CA as well as its corresponding slot-fillers at once. We utilize llama-2-70b-chat for both settings.

5.1 Zero-shot prompting results

For the evaluation of logic pattern template identification, we deem the final predicted pattern as correct if it is

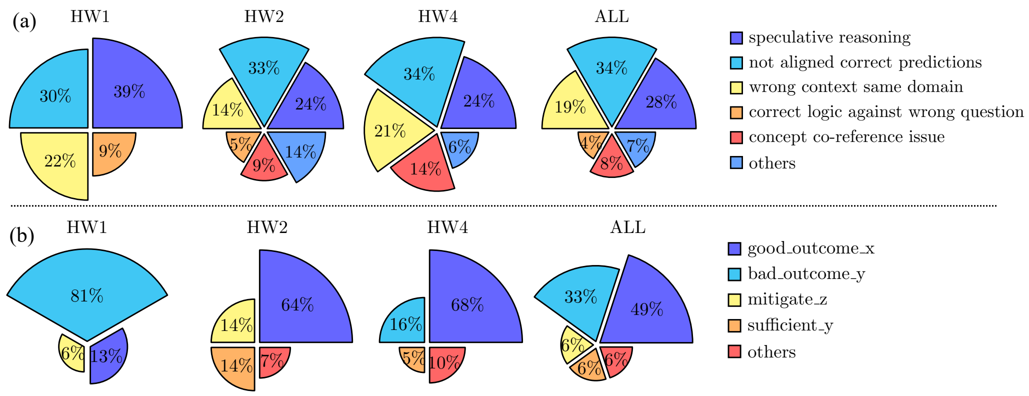


Figure 2 **(a)** Error types and their corresponding proportion. ***speculative reasoning***: model makes inferences to an extent that exceeds what can be explicitly inferred from the CA passage; ***not aligned correct predictions***: while the predicted pattern makes sense, it is not included in the list of agreed annotations; ***wrong context same domain***: model's explanation for the prediction does not refer to the actual context of the CA passage. Instead, it incorporates concepts from the same domain that are likely to co-occur with the question; ***correct logic against wrong question***: model's rationale for its prediction accurately elucidates CA's logic, however, the actual prediction is incorrect; ***concept co-reference issue***: model fails to relate the general concept presented in the question to the specific examples outlined in CA passage; ***others***: minor errors, including instances where the model fails to comprehend the CA passage due to its poor English, etc. **(b)** Root error-node distribution, indicating the proportion of specific nodes that act as the root cause for model's inaccurate predictions. Please refer to Figure 3 in the Appendix to see the actual questions for each node (ID).

present in the list of annotated patterns. For slot fillers, a manual evaluation is conducted due to the absence of an appropriate method for automatic evaluation. The predicted slot-filler is considered correct if the phrase aligns with the same meaning as one of the annotated slot-fillers. As shown in Table 2, our method consistently achieves superior precision scores compared to the baseline setting, excelling in both the identification of logic pattern templates and the extraction of slot fillers across all IAs.

5.2 Analysis

Toward further improving the model's performance, we conduct a comprehensive analysis on all predicted logic pattern templates that are not included in the annotations by investigating the explanation generated by the LLM for each question. Figure 2 illustrates the proportion of different error types and error nodes. Overall, the most dominant error is "speculative reasoning" which occurs when the LLM excessively infers information that is not explicitly stated in the given CA context. Moreover, the model provides most incorrect responses particularly when being queried about the presence of positive or negative outcomes in the CA essay. These two observations suggest that the questions associated with nodes "good_outcome_x" and "bad_outcome_y" are overly broad, lacking specificity to discern between similar templates. Consequently, this circumstance provides a space for excessive inference. To mitigate such issues, we intend to experiment with questions designed to distinguish similar templates more effectively. Additionally, one limitation in our approach is that a single error made at the upper levels of the decision tree would set the subsequent path astray, resulting in an incorrect final answer. To alleviate this issue, we intend to incorporate a checker component that assesses the decisions made at intermediate steps, and allows the LLM to backtrack to the previous step in future work.

6 Conclusion and Future Work

In this work, we explore addressing the intricate CA Logical Structure Analysis task by decomposing it with a decision tree. The experimental results underscore the efficacy of our approach. In addition to those mentioned in the Analysis 5.2, our future plans also include expanding the experimentation to encompass additional topics, allowing for comprehensive comparisons of results across different topics. Furthermore, although our current investigation solely comprises zero-shot experiments, given that our approach holds the advantage of being capable of autonomously generating training data without incurring human annotation costs since the series of answers to the identifiable questions on non-leaf nodes along the path to each logic pattern are unique, we plan to conduct few-shot learning and fine-tuning experiments to further test the efficiency of our method in the future.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 22H00524 and by JST CREST Grant Number JPMJCR20D2, Japan.

References

[1] Rosalind Driver, Paul Newton, and Jonathan Osborne. Establishing the norms of scientific argumentation in classrooms. **Science Education**, Vol. 84, pp. 1–312, 05 2000.

[2] 内藤昭一, 王文質, Paul Reisert, 井之上直也, Camélia Guerraoui, 山口健史, Jungmin Choi, Irfan Robbani, 乾健太郎. 反論の論理パターン解析: データセット構築と実現性検証. **30th Annual Meeting of the Natural Language Processing (NLP2024)**, 2024.

[3] Douglas Walton, Christopher Reed, and Fabrizio Macagno. **Argumentation Schemes**. Cambridge University Press, 2008.

[4] Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations, 2023.

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W Thomas, Florian Tramèr, Rose E Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. August 2021.

[6] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can't plan (a benchmark for LLMs on planning and reasoning about change). June 2022.

[7] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. December 2021.

[8] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey, 2023.

[9] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023.

[10] Jieyi Long. Large language model guided tree-of-thought, 2023.

[11] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In Jennifer Dy and Andreas Krause, editors, **Proceedings of the 35th International Conference on Machine Learning**, Vol. 80 of **Proceedings of Machine Learning Research**, pp. 2873–2882. PMLR, 10–15 Jul 2018.

[12] Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In **International Conference on Learning Representations**, 2020.

[13] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought prompting elicits reasoning in large language models. January 2022.

[14] Alon Talmor and Jonathan Berant. The Web as a Knowledge-Base for Answering Complex Questions. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, **Proceedings of the 2018 NAACL Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 641–651, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[15] Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. Multi-hop Reading Comprehension through Question Decomposition and Rescoring. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, **Proceedings of the 57th ACL**, pp. 6097–6109, Florence, Italy, July 2019. Association for Computational Linguistics.

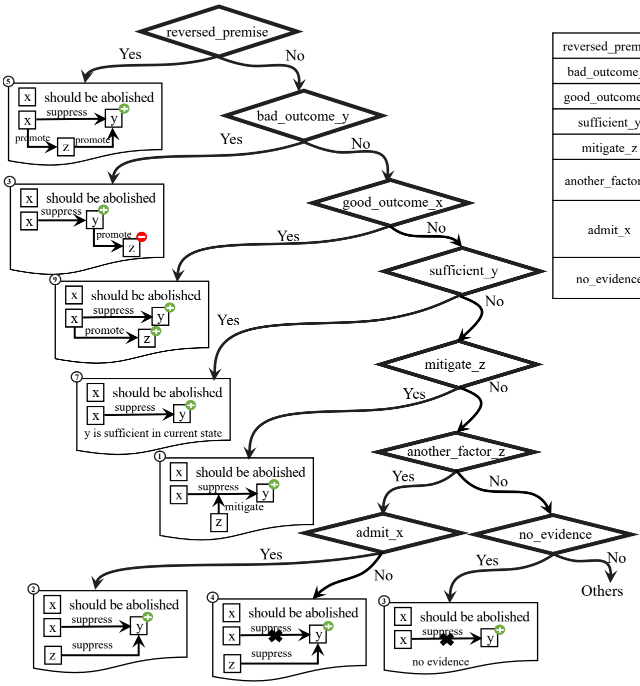
[16] Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, and Douwe Kiela. Unsupervised Question Decomposition for Question Answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, **Proceedings of the 2020 EMNLP**, pp. 8864–8880, Online, November 2020. Association for Computational Linguistics.

[17] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models, April 2023. arXiv:2205.10625 [cs].

[18] Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. Compositional Semantic Parsing with Large Language Models, September 2022. arXiv:2209.15003 [cs].

[19] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed Prompting: A Modular Approach for Solving Complex Tasks, April 2023. arXiv:2210.02406 [cs].

A Appendix



Question for each non-leaf node

reversed_premise	Does the counter-argument state that homework actually increases free time?
bad_outcome_y	Does the counter-argument state that free time causes a bad outcome?
good_outcome_x	Does the counter-argument mention any potential benefits of doing homework?
sufficient_y	Does the counter-argument state that free time is already sufficient in its current state?
mitigate_z	Does the counter-argument state that something mitigates 'homework reduces free time'?
another_factor_z	Does the counter-argument explicitly state that there is another factor besides homework that also reduces free time?
admit_x	Counter-argument explicitly states that there is another factor besides homework that reduces free time. Does the counter-argument also explicitly acknowledge that homework reduces free time?
no_evidence	Does the counter-argument explicitly state that the initial-argument does not provide any evidence to support the premise that 'homework reduces free time'?

Questions for querying "z" (slot-fiiler)

template①	Counter-argument acknowledges that homework reduces free time and that an action can be taken to increase free time. What is that action? The answer phrase should be extracted directly from the counter-argument.
template②	Counter-argument acknowledges that homework reduces free time but argues that there is another factor that also reduces free time, and it is more important to address that factor. What is that factor? The answer phrase should be extracted directly from the counter-argument.
template④	Counter-argument does not think that homework reduces free time. Instead, it argues that there is another factor that reduces free time. What is that other factor? The answer phrase should be extracted directly from the counter-argument.
template⑤	Counter-argument states that homework increases free time because homework promotes a thing that increases free time. What is that thing? The answer phrase should be extracted directly from the counter-argument.
template⑧	Counter-argument explicitly mentions a bad outcome which is directly caused by free time. What is that bad outcome? The answer phrase should be extracted directly from the counter-argument.
template⑨	Counter-argument mentions a potential benefit of doing homework. What is that benefit? The answer phrase should be extracted directly from the counter-argument.

Figure 3 The decision tree design and questions for querying LLMs for an IA which argues that "homework should be abolished since it reduces free time for hobbies, etc.".

Annotations	Baseline	Decomp
logic pattern templates: **[8]** slot-fillers: **{'ptn8': ['bad effect for students futures#not know where to start#concentrate', 'bad effect for students futures']}**	<s>[INST] <<SYS>> You are a master of logic. You can identify the logic pattern of a counter-argument in relation to an initial-argument. Answer the question regarding the logic pattern of counter-argument. <</SYS>> Identify the logical patterns of the following counter-argument in relation to the initial-argument. The logical pattern is one of the patterns in the following logical-pattern-descriptions. There might be multiple logical patterns within a counter-argument, but you only need to identify the most obvious one. Generate the identifier of the correct logical pattern and the corresponding {Z}. {Z} should be a phrase that is directly extracted from the counter-argument. [/INST] Understood! I'll do my best to help you with this task. Please go ahead and provide the initial-argument and counter-argument, and I'll start identifying the counter-argument's logical pattern.</s><s>[INST] Initial-argument: Today's topic is "Homework should be abolished". The point is "free time". We believe that if homework were to be abolished, we could have more free time. As a result, we could do more of what we really wanted like club activities, hobbies, or playing with friends. In my case, I go to tennis club after class until 5:00 pm and then I go to cram school until 8:00 pm. After this full day, I arrive at my home around 8:40 pm to eat dinner and take a shower. At nearly 10:00 pm I start my homework. I have a lot of homework. As a result, I go to bed late at night at nearly 1:00 am in the morning and I don't have the opportunity to sleep for a long period of time. It is not healthy. Therefore, homework should be abolished. Counter-argument: They said that we can get free time. However, it's not true. Because, giving freetime is bad effect for students futures. If you give too much free time to students, they may not know where to start when they go out into the world, which can be a problem. For example, in company we will doing a lot of task and we have to process everything. However, if we only take free time students will be concentrate in society. On this reason I think Homework shouldn't be abolish. Logical-pattern-descriptions: ptn1.The CA admits that homework reduces free time, but argues that the reduction of free time can be lessened by {Z} ptn2.The CA admits that homework reduces free time, but argues that there is another factor {Z} that also reduces free time, and it is more important to address {Z} ptn3.The CA asserts that there is no evidence to support the claim that homework reduces free time ptn4.The CA argues that homework does not reduce free time because there is another true factor {Z} that reduces free time ptn5.The CA argues that homework, on the contrary, promotes free time because it promotes the missed factor {Z} and {Z} promotes free time ptn6.The CA argues that homework, on the contrary, promotes free time because it suppresses the missed factor {Z} and {Z} suppresses free time ptn7.The CA argues that free time is not a desirable outcome because free time lacks value or is already sufficient and therefore does not require addressing ptn8.The CA argues that free time is not a desirable outcome because free time leads to a negative outcome {Z} ptn9.The CA argues that homework promotes a positive outcome {Z} from a different perspective than that of free time ptn10.The CA argues that homework mitigates a negative outcome {Z} from a different perspective than that of free time ptn11.The CA employs logic, but none of the aforementioned positions apply ptn100.No logical argument is presented, such as a greeting, introduction, or statement of stance What is the identifier of the logical pattern and its corresponding {Z}? [/INST] Output: ❌ **The logical pattern of the counter-argument is ptn7**, "The CA argues that free time is not a desirable outcome because free time lacks value or is already sufficient and therefore does not require addressing." The corresponding {Z} is "students' futures." The counter-argument suggests that giving students too much free time can have negative effects on their futures, as they may not know how to manage their time effectively or prioritize their responsibilities when they enter the workforce. Therefore, homework should not be abolished as it helps students develop important skills for their future.	########## Current prompt ############## Initial-argument: Today's topic is "Homework should be abolished". The point is "free time". We believe that if homework were to be abolished, we could have more free time. As a result, we could do more of what we really wanted like club activities, hobbies, or playing with friends. In my case, I go to tennis club after class until 5:00 pm and then I go to cram school until 8:00 pm. After this full day, I arrive at my home around 8:40 pm to eat dinner and take a shower. At nearly 10:00 pm I start my homework. I have a lot of homework. As a result, I go to bed late at night at nearly 1:00 am in the morning and I don't have the opportunity to sleep for a long period of time. It is not healthy. Therefore, homework should be abolished. Counter-argument: They said that we can get free time. However, it's not true. Because, giving freetime is bad effect for students futures. If you give too much free time to students, they may not know where to start when they go out into the world, which can be a problem. For example, in company we will doing a lot of task and we have to process everything. However, if we only take free time students will be concentrate in society. On this reason I think Homework shouldn't be abolish. The main point of the initial-argument is that homework reduces free time, thus homework should be abolished. Does the counter-argument state that homework actually increases free time? Please answer the question first, and then tell me your reason for your answer. Next, please tell me your confidence on your answer, your confidence should be represented as a decimal number between 0 and 1. Please output your confidence at a newline after a colon. LLM raw reply: No, the counter-argument does not state that homework increases free time. The counter-argument states that giving students too much free time can have negative effects on their future, such as lack of motivation and poor time management skills. My answer is based on the information provided in the counter-argument, which suggests that the speaker is concerned about the potential negative effects of excessive free time on students' future. Therefore, the speaker argues that homework should not be abolished. My confidence in this answer is 0.8. The reasoning behind my confidence is that the counter-argument directly addresses the initial-argument's point about free time, and provides a contrasting view that too much free time can be detrimental to students' future. Therefore, it is clear that the counter-argument does not support the idea that homework increases free time. ########## Current prompt ############## Does the counter-argument state that free time causes a bad outcome? Please answer the question first, and then tell me your reason for your answer. Next, please tell me your confidence on your answer, your confidence should be represented as a decimal number between 0 and 1. Please output your confidence at a newline after a colon. LLM raw reply: Yes, the counter-argument states that free time can cause a bad outcome. The reason for my answer is that the counter-argument mentions that giving students too much free time can be a problem, and suggests that it can lead to a lack of motivation and poor time management skills. This implies that the speaker believes that free time can have negative effects on students' behavior and performance. My confidence in this answer is 0.9. I am confident in my answer because the counter-argument explicitly states that free time can have negative effects on students' future, and suggests that homework can help to mitigate these effects. This implies that the speaker believes that free time can cause a bad outcome for students. **predicted_ptn: ptn8** ✅ ########## Current prompt ############## Counter-argument explicitly mentions a bad outcome which is directly caused by free time. What is that bad outcome? The answer phrase should be extracted directly from the counter-argument. Please extract the exact phrase from counter-argument without modifying and write down your answer after a colon: Please also give me your confidence score on this question as well at a newline after a colon. LLM raw reply: The bad outcome mentioned in the counter-argument is: "**they may not know where to start when they go out into the world, which can be a problem**." My confidence in this answer is 1.0. This phrase is mentioned explicitly in the counter-argument as a potential negative consequence of giving students too much free time. It suggests that without the structure and responsibility of homework, students may struggle with independent task management and decision-making, which can lead to difficulties in their future careers and personal lives.

Figure 4 An example of the actual prompt and model's generation for both **Baseline** and **Decomp** settings.