DIMSUM: Discourse in Mathematical Reasoning as a Supervision Module

Krish Sharma[1,2,†], Niyar R Barman[1,2,†], Akshay Chaturvedi[1], Nicholas Asher[1]

[1]IRIT, Toulouse, France
[2] NIT Silchar, India

Abstract

We look at reasoning on GSM8k, a dataset of short texts presenting primary school, math problems. We find, with Mirzadeh et al. (2024), that current LLM progress on the data set may not be explained by better reasoning but by exposure to a broader pretraining data distribution. We then introduce a novel information source for helping models with less data or inferior training reason better: discourse structure. We show that discourse structure improves performance for models like Llama2 13b by up to 160%. Even for models that have most likely memorized the data set, adding discourse structural information to the model still improves predictions and dramatically improves large model performance on out of distribution examples.

1 Introduction

Recent advancements in large language models (LLMs) have led to impressive performance on reasoning tasks, particularly on benchmark datasets like GSM8K (Cobbe et al., 2021). These models exhibit strong problem-solving abilities, often producing solutions that resemble human-like reasoning. However, recent studies have demonstrated that altering the entities or numerical values can degrade the reasoning capabilities of LLMs (Mirzadeh et al., 2024). Other studies have also shown that model reasoning is not robust (Schaeffer et al., 2024; Valmeekam et al., 2022; Asher and Bhar, 2024; Kambhampati, 2024) . This raises a critical question: do these models genuinely engage in reasoning, or do they primarily rely on memorization and pattern recognition? These findings suggest a need for a deeper investigation into how LLMs process logical structures and linguistic variations.

Prior research (Chen et al., 2024) has also investigated the effect of permuting the premise order in GSM8K problems, showing that such modifications affect LLMs' reasoning capabilities. This observation suggests that, unlike formal logic, where the validity of a conclusion remains unchanged regardless of premise order, natural language reasoning is shaped by structural dependencies within discourse.

A linguistic perspective explains this dependency. Unlike formal logic, natural language introduces ambiguity and underspecification, which are resolved through contextual and structural cues that tell us how clauses in a text are semantically related. The ordering of clauses within a discourse, for instance, significantly impacts interpretation, and determines, in the absence of other cues, temporal and causal relationships. Consider the simple sequences in (1) from (Asher, 1993):

(1) a. Pat took off his shoes. Pat got in bed.
  b. Pat got in bed. Pat took off his shoes

When we permute the order of the clauses as in (1)b, most native English speakers will draw different conclusions. They will conclude from in (1)a that Pat first took off his shoes before getting into bed, whereas in (1)b that Pat took his shoes off after he got in bed. Ordering in this case determines a semantic relation–(Asher, 1993) calls it Narration that entails that the event in the second clause follows the event in the first.

Even in seemingly simple narratives, semantic relations such as Elaboration, Background information, and Narration shape how reasoning unfolds. Additionally, discourse structure affects anaphora resolution—reordering premises can change how pronouns and referents are interpreted, thereby influencing reasoning outcomes. For example, in (2)

(2) John picked 3 apples. Sam picked 4 apples. He then picked 2 more.

[†]Equal Contribution

Proceedings of the 26th Annual Meeting of SIGDIAL, pages 294–310
Aug 25–27, 2025. ©2025 SIGDIAL

most native speakers would pick Sam as the referent of *he* in the last sentence. If we reverse the first two sentences, most speakers would pick John as the referent of *he*.

We show that information about discourse structure improves performance for all LLMs on the GSM8k data set, but especially for older, models like Llama2 13b with poor reasoning performance by up to 160%. Even for models that have most likely memorized the data set, adding discourse structural information to the model improves performance on GSM8K and dramatically improves performance on out of distribution examples.

2 Hard GSM8K Dataset and Variants

To systematically assess whether LLMs really reason or just rely on memorized patterns, we use the more difficult portion of the GSM8K data set, *Hard GSM8K*, (with problem statements involving 4 or more premises). We also provide a controlled variant of Hard GSM8K to create out-of-distribution (OOD) variants using the transformations defined below (see Appendix A.5 for examples). This data will be made publicly available. Unlike the data in Mirzadeh et al. (2024), where modifications involves changing entity names and adding or subtracting premises, our variants are more complex in terms of contextual modification and number of substitutions. We investigate whether the model can maintain performance under variations in numerical values and contextual framing. This data enables us to rigorously evaluate the robustness of LLMs' reasoning, and it serves as a test-bed to study the influence of the discourse structure.

2.1 Contextual Modification (C-MOD)

In this transformation, we alter the real-world context of a problem while preserving its structural framework and numerical relationships, ensuring the modified problem shifts out of the pretraining data distribution. Given a problem statement P, we define a contextual mapping function $f_C : P \to P'$, where entities and actions are replaced by semantically distinct counterparts from an unrelated domain. Formally, if $E = \{e_1, e_2, \ldots, e_n\}$ represents entities in the original problem and $A = \{a_1, a_2, \ldots, a_m\}$ denotes actions, then:

$$f_C(E, A) = (E', A') \quad \text{such that} \quad \begin{array}{l} E' \cap E = \emptyset, \\ A' \cap A = \emptyset. \end{array} \tag{1}$$

For instance food items are substituted with digital files, and weight was mapped to storage size.

2.2 Numerical Modification (N-MOD)

This transformation alters numerical values while preserving the problem's reasoning structure. Given a set of numerical values $N = \{n_1, n_2, \ldots, n_k\}$ in a problem, we apply a function $f_N : N \to N'$ such that:

$$\forall n_i \in N, \quad f_N(n_i) = n_i', \\ \text{where} \quad \text{Scale}(N') = \text{Scale}(N).$$

Here, $\text{Scale}(N)$ ensures proportionality is maintained. For instance, if the weight of an object triples, the same transformation is applied to the modified numbers.

2.3 Lexical Modification (L-MOD)

This technique modifies words and phrases without altering sentence structure or meaning. Given a vocabulary set $V = \{v_1, v_2, \ldots, v_p\}$, we define a substitution function $f_L : V \to V'$ such that:

$$\forall v_i \in V, \quad f_L(v_i) = v_i', \quad \text{where } \text{Syn}(v_i) = v_i'.$$

Here, $\text{Syn}(v)$ ensures v' is a valid synonym or equivalent phrase. This tests whether models are robust to surface-level changes.

3 Generating discourse structure

To annotate discourse structures systematically within Hard GSM8K, we identify 10 complex examples containing multiple events and subjects, with more than eight sentences. We manually annotated these examples according to rules in (Asher and Lascarides, 2003), used them as few-shot (Brown et al., 2020) exemplars for generating annotations across the entire dataset using the Llama 3.1 70B model.

3.1 Generalized Rules for Annotating Discourse Structure

Predicting an SDRT discourse structure (SDRS) of a text or conversation requires a series of steps (Asher et al., 2016; Bennis et al., 2023; Thompson et al., 2024) that we follow here: (i) identifying the basic or elementary discourse units (EDUs),

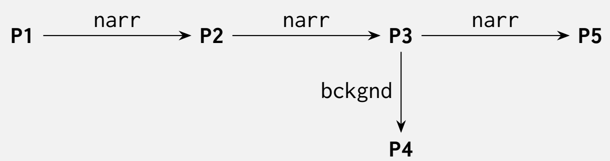
EXAMPLE OF DISCOURSE STRUCTURE:

Question: Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

Discourse Structure:
<Structure>
Topics:
[topic-a] Oliver's kiwi picking on consecutive days
[topic-b] Sunday's special condition (five smaller kiwis)

Relationships: $t_a \xrightarrow{\text{Elab}} t_b$
Premises:
P1: Picks 44 kiwis on Friday P2: Picks 58 on Saturday P3: On Sunday picks double the Friday amount P4: Sunday has five smaller kiwis P5: Oliver wants to know total kiwis

Narrative Flow:



</Structure>

Figure 1: Overview of the prompt sequence on an example from (Mirzadeh et al., 2024) that GPT-o1-mini and Llama-3-8B couldn't solve. With this prompt all models tested solved the problem correctly (see Appendix A.4.2 for model generation outputs). The full structure generation prompt (A.3.1), answer generation prompt (A.3.2) and few-shot examples (A.4) are available in the appendix .

each of which contains one main event, state or concept; (ii) attaching the EDUs together to form an unlabelled graph; (iii) labeling the edges of the graph with semantic relations. In math problem corpora, the premises typically have simple clausal structures, and so each premise contributes one EDU to the SDRS graph. The labelled edges in the graph give the semantic relationships between the premises.

Following the instructions in the SDRT annotation manual for the STAC corpus[1], we offer a set of informal rules in the system prompt to help models build a discourse structure, incorporating temporal markers as a key factor. These informal rules capture the key temporal and thematic aspects of the semantics of the SDRT relations that are relevant here Asher and Lascarides (2003).

3.2 Identifying Eventualities

Each premise in our data set is a sentence or clause; and in a discourse structure each one introduces a central *eventuality*, an event, state or concept (Asher and Lascarides, 2003). The appendix provides some guidelines in this regard A.1 - E_{P1}: kiwis picked on Friday. - E_{P2}: those picked on Saturday. - E_{P3}: kiwis picked on Sunday - E_{P4}: smaller kiwis on Sunday - E_{P5}: how many kiwis in total.

[1] https://www.irit.fr/STAC/stac-annotation-manual.pdf

3.3 Relations Between Clauses and eventualities

Once eventualities are identified, their relationships are fixed by various discourse relations like Elaboration, Narration or Background that hold between the premises that express them. The following glosses help specify the meanings of these relations and how to infer them.

Narration (Narr(P_i, P_j)**)**– When P_j introduces an eventuality that sequentially follows the one introduced by P_i. Narration can be established with explicit temporal markers in P_i and P_j (*then , next, after that, subsequently, finally,* or explicit sequentially related times like consecutive days of the week or months of the year) or when P_i and P_j form part of a chain of actions leading to the final computation.

Model	GSM-Symbolic						GSM-MOD (Ours)							
	Hd-GSM8K		P1		P2		C-MOD		N-MOD		L-MOD		Overall	
	DS-	DS+	DS-	DS+	DS-	DS+	DS-	DS+	DS-	DS+	DS-	DS+	DS-	DS+
Llama 3.1-8B	79.2	88.2	58.2	79.4	34.6	70.2	32.3	71.7	73.7	77.8	56.6	75.7	54.2	75.1
Llama 3-8B	70.2	79.8	52.8	73.6	28.3	62.4	24.2	62.6	64.6	71.7	52.5	69.7	47.1	68.0
Llama 2-13B	18.4	48.2	8.2	34.4	6.8	34.2	9.1	39.4	12.1	41.4	10.1	42.4	10.4	41.1
Llama 3.1-70B	96.0	98.3	86.8	93.3	85.1	90.2	59.6	89.9	94.9	97.0	80.8	91.9	78.4	92.9

Table 1: Accuracies comparison of different Llama models on Hard GSM8K, Mirzadeh et al. (2024)'s P1 and P2 datasets and our perturbation variants (C-MOD, N-MOD, L-MOD) . DS- = without discourse structure ; DS+ = with discourse structure. Overall score siginfies average of all variants of GSM-MOD. The models with discourse structure (see Figure 1) perform considerably better than their counterpart.

Elaboration (Elab(P_i, P_j)**)**– When P_j provides additional details about the event in P_i, making it clearer or more informative. Elaboration also occurs when P_j introduces an intermediate computation step necessary for understanding P_i.

Background (Bckgd(P_i, P_j)**)**– When P_j provides contextual information about P_i that is not directly required for computation but helps in understanding the problem. This is commonly the case when P_j describes an event that happened before P_i but does not impact the reasoning process. Markers such as on *Monday, previously, before, in the past* indicate background relationships. Shift from an active verb sequence to a stative verb (underlined in Figure 1) can also signal Background. For example, consider the problem statement from Figure 1; P4 is attached with Background to P3, and P5 attaches to P3 with Narration. The attachment reinforces the constraint that the background information is not to be used in the main computation.

3.4 Markers for positions in a narrative

Discourse structure also exploits markers that tell us about the order of clauses in a text. The simplest consists of an ordinal enumerations as in *first, P1, second P2, third P2, fourth P4*. In addition, there are markers for first and last clauses: *initially, finally*. When clauses containing these markers appear in positions that contradict their intended sequence, it becomes necessary to reorder the premises to align with the canonical narrative flow. We apply this principle to the examples in Chen et al. (2024) (unfortunately they did not release the whole dataset).

4 Results

We evaluated several Llama models (Touvron et al. (2023) Grattafiori et al. (2024)) –Llama 3.1 70B, Llama 3.1 8B, Llama 3 8B, and Llama 2 13B–on reasoning with three distinct data sets, Hard GSM8K and variants, GSM symbolic and our variants and our GSM Mod. We tested the performance of models on problems with and without incorporated discourse structural information. We used a four-shot learning setting. We also tested our approach on Mirzadeh et al. (2024)'s data sets with additional premises that were needed (their P1 and P2). To replicate their method, we used eight shot learning. Table 1 shows our results.

Table 1 clearly shows that when models have access to the discourse structure of the problems, their performance improves substantially on the original GSM8K problems, on the variants that we have constructed, and on Mirzadeh et al. (2024)'s data sets. This held true both for large models, for which it is pretty clear that they have seen GSM8k problems in their training, and for older smaller models, whose performance drastically improved with access to discourse structure. Adding discourse structure allows the poorest performing models catch up somewhat to the better ones. Table 1 also shows that training regimes affect predictions; Llama 3 and 3.1 share largely the same architecture but have quite different training regimes and make significantly different predictions.

Even more noteworthy is the fact that on our altered, unseen examples, even large models like Llama3.1 70b saw their prediction success rate drop, by almost 40 percentage points for the contextually shifted stories. Once those models had access to the discourse structure their performance improved by sometimes up to 30 percentage points. From the table, it is clear that the contextually shifted stories were much more challenging than the other variants we made.

We also tested models on examples with background information from Mirzadeh et al. (2024), and all models correctly avoided using the background information in the computation. In Appendix A.4.3, we also show that the Llama 3.1 70B with discourse structure are able to predict the correct answers to all the examples provided in Chen et al. (2024).

Finally, we tested our models on a subset of the (Mitra et al., 2024). This is a synthetically generated data set for training, not a benchmark. But we felt that evaluating the usefulness of discourse structure on this data set might show whether our approach overfitted the GSM8K paradigm. We used an 8 shot learning regime, an observed a very significant boost in scores as in Table 2 on this data set when the model had access to the generated discourse structure of the problems in this data set.

Model	Configuration	Accuracy (%)
LLaMA 3.1 70B	DS-	75
LLaMA 3.1 70B	DS+	81
LLaMA 3.1 8B	DS-	58
LLaMA 3.1 8B	DS+	69

Table 2: Accuracy comparison between base (DS-) and DIMSUM (DS+) versions of LLaMA 3.1 models.

5 Conclusion

We have shown that providing discourse structure, which encodes the semantic relations between premises or clauses in a text, significantly improves the performance of language models on tasks such as those in the GSM8K dataset. Our results also demonstrate that even models that perform well on the standard dataset show substantially degraded performance on out-of-distribution variants; however, incorporating discourse structure leads to notable gains in robustness and generalization across these variants.

As a future direction, we recognize the importance of scaling the generation of discourse structures. While our current approach relies on LLaMA 3.1 70B, exploring the feasibility and effectiveness of generating discourse structures using smaller language models is a promising avenue. This would broaden the accessibility and applicability of our method, enabling lightweight integration in resource-constrained settings.

Limitations

One limitation of our paper is that the smaller models do not generate the discourse structure themselves. Llama70b provides the discourse structure. We think that models in the 8B range would need fine tuning to learn such structures (Thompson et al., 2024). We plan to use Llama70b as an annotator so that the smaller models can be fine tuned to provide this structure on problem sets like GSM8k.

We did not test our system on GSM-IC (Shi et al., 2023), because upon inspection many of the problems in the data set involve the addition of *irrelevant* information. Consider this example from GSM-IC for instance.

(3) Lucy has $65 in the bank. She made a $15 deposit and then followed by a $4 withdrawal. The shoe size of Lucy's brother is 80. What is Lucy's bank balance?

The brother's shoe size really is irrelevant to anything in the story, and including it actually makes the text kind of incoherent at least very awkward. From a discourse structure perspective, the sentence about shoe size intuitively isn't attached to any other clause in the story. Background is different; Background information is relevant but it is like stage setting or additional information about some object or event that is part of the main narrative. Our model is trained to find Backgrounds but not irrelevant information.

Acknowledgments

We gratefully acknowledge the support of the AI Cluster ANITI (ANR-19-PI3A-0004) and the Rakel grant. This work used the HPC resources from CALMIP (Grant 2016-P23060) and TURPAN (Grant M24047). We also acknowledge the support of IRT Saint Exupéry through their AIRSTRIP project call of November 2023: *AI in the service of aeronautical and spatial systems*.

Ethics Statement

We have demonstrated that leveraging discourse structure can improve performance, although it remains imperfect. This work is grounded in mathematical reasoning, which is a prerequisite for downstream real-world applications involving human–robot collaboration. A lack of numerical understanding in a model can lead to tangible consequences, as humans often rely on numerical expressions when providing instructions to robots. Our findings with variants of the gsm-mod model indicate that these models may rely heavily on memorized data, as further evidenced by gsm-symbolic. Consequently, their reported performance metrics may be inflated, potentially creating a misleading sense of readiness for deployment in real-world scenarios.

References

Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers.

Nicholas Asher and Swarnadeep Bhar. 2024. Strong hallucinations from negation and how to fix them. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12670–12687.

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: The STAC corpus. In *Language Resources and Evaluation Conference*, pages 2721–2727.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, New York, NY.

Zineb Bennis, Julie Hunter, and Nicholas Asher. 2023. A simple but effective model for attachment in discourse parsing with multi-task learning for relation labeling. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3404–3409.

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. *Preprint*, arXiv:2005.14165.

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. *Preprint*, arXiv:2402.08939.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas

Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Subbarao Kambhampati. 2024. Can large language models reason and plan? *Annals of the New York Academy of Sciences*, 1534(1):15–18.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024. Orca-math: Unlocking the potential of slms in grade school math. *Preprint*, arXiv:2402.14830.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2024. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. *arXiv preprint arXiv:2302.00093*.

Kate Thompson, Akshay Chaturvedi, Julie Hunter, and Nicholas Asher. 2024. Llamipa: An incremental discourse parser. *arXiv preprint arXiv:2406.18256*. In *EMNLP 2024 Findings*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). In *NeurIPS 2022 Foundation Models for Decision Making Workshop*.

A Appendix

In this appendix, we provide additional details to the main text, including:

- A.1 Guidelines for Finding Eventualities
- A.2 Experimental Setup
- A.3 System Prompts
- A.4 Examples of Answer Generation
- A.5 GSM-MOD Example

A.1 Guidelines for Finding Eventualities

- **Main Subject:** The entity or concept driving the problem statement (e.g., a person making a purchase, a worker repairing equipment).
- **Key Actions:** Events that modify the subject's state (e.g., buying, losing, doubling, converting).
- **Transitions Between Events:** Phrases like *then, after that, subsequently* indicate topic shifts.
- **Supporting Context:** Background information that does not directly influence the problem's computation.

A.2 Experimental Setup

All experiments were conducted at the inference stage using an A100 GPU as the primary computing resource. Each experiment was performed as a single run, utilizing approximately 600 GPU hours. The experiments were implemented using the `transformers` library.

The hyperparameters and settings for answer generation are as follows:

```
max_new_tokens: 400
temperature: 0.4
top_p: 0.9
tokenizer.pad_token_id: 18610
tokenizer.padding_side: 'right'
```

A.3 System Prompts

The system prompts used to generate both the discourse structure (A.3.1) and the final answers (A.3.2) to each query are below.

A.3.1 Prompt for structure generation

```
Task: Analyze a short story using a structured relational framework,ensuring proper sequencing and relational mapping.
  Instructions:
  Identify Topics and Premises: - Assign meaningful topic labels to key elements of the story (e.g., [topic-a]: Initial context, [topic-b]: Character's key action). - List premises (P1, P2, ..., Pn) capturing essential events or actions.
  Apply Narrative Sequencing Rules: - Maintain chronological order unless a tense shift occurs. - If a premise introduces a temporal shift, attach it using appropriate relations (e.g., Narr for sequential events, Bkg for background details). - Adjust premise order based on discourse markers and time references.
  Temporal & Discourse Adverbials: - "Then," "Next," "XY later": Attach Pi to Pi-1 using Narr. - "XY before": Attach Pi to Pi-1 using Narr for reverse ordering. - "While": Use Elaboration (elab) to connect related events. - "Finally": Ensure Pi is the last premise. - "First": Ensure Pi is the initial premise.
  Day-Specific Rules: - If Pi: "Day n B" and Pi+1: "Day n+1 B," attach Pi to Pi+1 using Narr. - If days are out of order, shuffle until proper sequential flow is restored.
  Label Relationships Between Elements: - Use t1-relation-t2 for topic-level relationships (e.g., cause-effect). - Use PX-relation-PY for premise-level relationships.
  Output Format:
  Topics: - List identified topics with brief descriptions.
  Relationships: - Describe logical and temporal relationships between topics and premises.
  Premises: - Present premises (P1, P2, ..., Pn) in a logically ordered sequence.
  Narrative Structure: - Show premise connections based on sequencing rules.
  Exclusions: - Do not include resolution, calculations, or final answers.
```

A.3.2 Prompt for answer generation

Task: Given a set of premises and an abstract that defines their properties, determine the answer to the question using only the information provided in the abstract. The abstract provides a structured relational framework, ensuring logical consistency in reasoning.

Instructions:

Identify Topics and Premises: - Assign meaningful topic labels to the key elements of the problem (e.g., [topic-a]: Initial context, [topic-b]: Events influencing the outcome). - List the premises (P1, P2, ..., Pn), ensuring they contain all necessary descriptive statements.

Apply Narrative Sequencing Rules Between Premises: - Maintain chronological order unless explicitly defined otherwise in the abstract. - If a premise introduces a temporal or logical shift, attach it accordingly using appropriate relations (e.g., Narr for sequential events, Bkg for background information). - Resolve premise reordering constraints based on given discourse markers.

Use the Abstract's Structured Information to Derive the Answer: - Follow the relationships and premises as structured in the abstract. - Do not introduce external knowledge; rely solely on the provided abstract and premises. - If numerical reasoning is required, apply logical derivations based on the premises.

Output Format:

<answer> - Provide the final computed or inferred answer. - Ensure it aligns with the logical structure in the abstract. </answer>

A.4 Examples of Answer Generation

A.4.1 GSM-MOD

C-MOD:

Original Question: Ken created a care package to send to his brother, who was away at boarding school. Ken placed a box on a scale, and then he poured into the box enough jelly beans to bring the weight to 2 pounds. Then, he added enough brownies to cause the weight to triple. Next, he added to the package gummy bears whose weight equaled 1/2 of the weight of the package with the brownies and the jelly beans. And finally, he added enough gummy worms to double the weight once again. What was the final weight of the box of goodies, in pounds?

Modified Question: Ken created a package to send to his brother help him repair his house. Ken placed a box on a scale, and then he poured into the box enough nails to bring the weight to 2 pounds. Then, he added enough screws and bolts to cause the weight to triple. Next, he added to the package a set of socket wrenches whose total weight equaled 1/2 of the weight of the package so far. And finally, he added a bunch of drills and bits to double the weight once again.

Structure:

<structure>
Topics [topic-a]: Ken's package creation for his brother's house repairs [topic-b]: Incremental weight additions (nails, screws, bolts, socket wrenches, drills)
Relationships t1–elab–t2
Premises P1: Ken created a package to send to his brother to help repair his house. P2: He placed a box on a scale and poured in nails until the weight reached 2 pounds. P3: He then added enough screws and bolts to cause the weight to triple. P4: Next, he added a set of socket wrenches weighing half the package's current total. P5: Finally, he added drills and bits to double the weight once again.
Narrative Structure P1–narr–P2 P2–narr–P3 P3–narr–P4 P4–narr–P5
</structure>

Answer without DS (generated by Llama 3.1 70B):

Step 1: Determine the initial weight of the nails. The initial weight of the nails is 2 pounds.
Step 2: Calculate the weight after adding screws and bolts. After adding screws and bolts, the weight triples. So, the new weight is 2 * 3 = 6 pounds.
Step 3: Calculate the weight after adding the socket wrenches. The weight of the socket wrenches is 1/2 of the current

weight. So, the weight of the socket wrenches is 6 / 2 = 3 pounds. The total weight now is 6 + 3 = 9 pounds.
Step 4: Calculate the weight after adding the drills and bits. The weight doubles after adding the drills and bits. So, the new weight is 9 * 2 = 18 pounds. However, we know the final weight is 42 pounds, not 18 pounds, which means our calculation approach needs adjustment to find the weight of the drills and bits
Step 5: Correct the approach to find the weight of the drills and bits. Given that the final weight of the package is 42 pounds and this is after the weight has doubled from the addition of the drills and bits, we can work backwards. If the weight doubles to 42 pounds, then before adding the drills and bits, the weight was 42 / 2 = 21 pounds.
Step 6: Calculate the weight of the drills and bits. Since the weight before adding the drills and bits was 21 pounds, and after adding them it became 42 pounds, the weight of the drills and bits is 42 - 21 = 21 pounds.
The final answer is: 21

Answer with DS (generated by Llama 3.1 70B)
Step 1: Identify the initial weight of the package The initial weight of the package after adding nails is 2 pounds, as stated in premise P2.
Step 2: Calculate the weight after adding screws and bolts The weight triples after adding screws and bolts, so the new weight is 2 * 3 = 6 pounds, as stated in premise P3.
Step 3: Calculate the weight after adding socket wrenches The weight of the socket wrenches is 1/2 of the current weight of the package, which is 6 pounds. So, the weight of the socket wrenches is 6 / 2 = 3 pounds. The total weight now is 6 + 3 = 9 pounds, as stated in premise P4.
Step 4: Calculate the final weight after adding drills and bits The weight doubles after adding drills and bits, so the final weight is 9 * 2 = 18 pounds, as stated in premise P5.
The final answer is: 18

N-MOD:

Original Question: Sanjay saw a 40-foot dolphin with 12 16-inch remoras attached to it. But a quarter of the remoras go away. What percentage of the dolphin's body length is the combined length of the remaining remoras?

Modified Question: Sanjay saw a 60-foot dolphin with 16 12-inch remoras attached to it. But a quarter of the remoras go away. What percentage of the dolphin's body length is the combined length of the remaining remoras?

Structure:
<structure>
Topics [topic-a]: Sanjay's observation of the dolphin [topic-b]: The remoras attached to the dolphin
Relationships t1-elab-t2
Premises P1: Sanjay saw a 60-foot dolphin with 16 12-inch remoras attached to it. P2: But a quarter of the remoras go away.
Narrative Structure P1-narr-P2
</structure>

Answer without DS: First, find the total number of remoras remaining: 16 - 16*1/4 = 12. Then, find the combined length of the remoras in inches: 12 inches/remora * 12 remoras = 144 inches. Then divide that number by 12 to convert it to feet: 192 inches / 12 inches/foot = 16 foot. Then divide the combined remora length in feet by the dolphin's length and multiply by 100% to express the answer as a percentage: 16 foot / 60 feet * 100% = 26%

Answer with DS: <answer>

From P1, we know that Sanjay saw a 60-foot dolphin with 16 12-inch remoras attached to it. From P2, we know that a quarter of the remoras go away.

Let's first calculate the length of the remaining remoras. If a quarter of the remoras go away, then the number of remoras remaining is:

16 - (16 / 4) = 16 - 4 = 12

So, there are 12 remaining remoras. Their combined length is:

12 * 12 = 144 inches

Now, let's calculate the percentage of the dolphin's body length that the combined length of the remaining remoras is. We know the dolphin is 60 feet long, which is equal to 720 inches (60 * 12).

The percentage of the dolphin's body length that the combined length of the remaining remoras is:

(144 / 720) * 100

Therefore, the combined length of the remaining remoras is 20% of the dolphin's body length.

</answer>

A.4.2 GSM-Symbolic-NoOp

The example in this subsection is taken from the Symbolic-NoOP variation of (Mirzadeh et al., 2024), where strong reasoning models like o1-mini(OpenAI et al., 2024) fail without discourse structure. However, with the correct discourse structure, all models in the Llama family, including Llama 2 13B, produce the correct answer.The highlighted text in answer generation demonstrates how it corrects its reasoning. This demonstrates that with an appropriate discourse structure representation, smaller models can be more robust than state-of-the-art models.

Question: Oliver picks 44 kiwis on Friday. Then he picks 58 kiwis on Saturday. On Sunday, he picks double the number of kiwis he did on Friday, but five of them were a bit smaller than average. How many kiwis does Oliver have?

Structure:

<structure>

Topics [topic-a]: Oliver's kiwi picking activities on consecutive days [topic-b]: Sunday's special condition (five smaller kiwis)

Relationships t1-elab-t2

Premises P1: Oliver picks 44 kiwis on Friday. P2: Then he picks 58 kiwis on Saturday. P3: On Sunday, he picks double the number of kiwis he did on Friday. P4: Among Sunday's kiwis, five are smaller than average. P5: Oliver wants to know the total number of kiwis he has.

Narrative Structure P1-narr-P2 P2-narr-P3 P3-bckgnd-P4 P3-narr-P5

</structure>

Answer without DS (generated by o1-mini as per (Mirzadeh et al., 2024)):

Oliver picks kiwis over three days as follows: - Friday: 44 kiwis - Saturday: 58 kiwis - Sunday: Double the number he picked on Friday, which is 2 × 44 = 88 kiwis However, on Sunday, 5 of these kiwis were smaller than average. We need to subtract them from the Sunday total: 88 (Sunday's kiwis) - 5 (smaller kiwis) = 83 kiwis Now, summing up the kiwis from all three days: 44 (Friday) + 58 (Saturday) + 83 (Sunday) = 185 kiwis. Oliver has a total of 185 kiwis.

Answer with DS (generated by Llama-3.1-8B):

<answer> To find the total number of kiwis Oliver has, we need to follow the premises as structured in the abstract.

From P1, we know Oliver picks 44 kiwis on Friday.

From P2, we know he picks 58 kiwis on Saturday.

From P3, we know on Sunday, he picks double the number of kiwis he did on Friday. Double the number of kiwis from Friday is 2 * 44 = 88 kiwis.

However, P3 is connected to P4 as background information, which means means the information about Sunday's kiwis is additional context, but it doesn't affect the count of kiwis.

From P5, we know Oliver wants to know the total number of kiwis he has.

To find the total number of kiwis, we add the kiwis picked on Friday, Saturday, and Sunday. Total kiwis = 44 (Friday) + 58 (Saturday) + 88 (Sunday) = 190 kiwis.

The five smaller kiwis mentioned in P4 do not affect the total count, as they are a subset of the kiwis picked on Sunday.

Therefore, Oliver has 190 kiwis. </answer>

A.4.3 R-GSM

The examples of this subsection is taken from the R-GSM examples mentioned in (Chen et al., 2024).We see that GPT4-turbo fails to understand the semantics of the question and evaluate the order of sequences whereas Llama 70B with discourse structure get it.

Question: Thomas withdraws $1000 in 20 dollar bills from the bank account. After getting home, he uses half of the remaining bills to pay for a bill. Thomas then triples his money. He then converts all his bills to 5 dollar bills. He lost 10 bills while getting home. How many 5 dollar bills does he have?

Correct Answer: 240

Structure:

<structure>

Topics [topic-a]: Thomas's initial withdrawal [topic-b]: Subsequent money actions
Relationships t1–elab–t2
Premises P1: Thomas withdraws $1000 in 20 dollar bills from the bank account. P2: He lost 10 bills while getting home. P3: After getting home, he uses half of the remaining bills to pay for a bill. P4: Thomas then triples his money. P5: He then converts all his bills to 5 dollar bills.
Narrative Structure P1–elab–P2 P2–narr–P3 P3–narr–P4 P4–narr–P5
</structure>

Answer without DS: (generated by GPT-4-Turbo):

1. Thomas withdraws $1000 in $20 bills from the bank. - $1000 / $20 = 50 bills
2. He uses half of the remaining bills to pay for a bill. - Half of 50 bills is 25 bills - So, he uses 25 bills and has 25 bills left
3. Thomas then triples his money. - He had 25 bills of $20 each - 25 * $20 = $500 before tripling - Tripling $500 gives him $500 * 3 = $1500
4. He then converts all his bills to $5 bills. - $1500 / $5 = 300 bills
5. He lost 10 bills while getting home. - 300 bills - 10 bills = 290 bills

Final Answer: Thomas has 290 $5 bills.

Answer with DS:

<answer> To determine the number of 5 dollar bills Thomas has, we will follow the premises in the given order.

1. Thomas withdraws $1000 in 20 dollar bills from the bank account (P1). Since each bill is $20, the number of bills is $1000 / $20 = 50 bills.
2. He lost 10 bills while getting home (P2). The remaining number of bills is 50 - 10 = 40 bills.
3. After getting home, he uses half of the remaining bills to pay for a bill (P3). Half of the remaining bills is 1/2 * 40 = 20 bills. He uses 20 bills to pay for a bill, leaving him with 40 - 20 = 20 bills. Since each bill is $20, the total amount of money he has is 20 * $20 = $400.
4. Thomas then triples his money (P4). He triples the $400, resulting in 3 * $400 = $1200.
5. He then converts all his bills to 5 dollar bills (P5). Since he now has $1200 and each bill is $5, the number of 5 dollar bills is $1200 / $5 = 240 bills.

Therefore, Thomas has 240 5 dollar bills. </answer>

Question: Larry loves taking care of animals. He has 3 cats. He has 3 times as many dogs as cats. He has 2 fewer rabbits than dogs. He also has a collection of gerbils that's 1/3 the number of fish he has. He has a fish tank with three times the number of fish as rabbits. How many pets does Larry have?

Correct Answer: 47

Structure:

<structure>
Topics [topic-a]: Larry's interest in animals [topic-b]: Inventory of Larry's pets
Relationships t1–elab–t2
Premises P1: Larry loves taking care of animals. P2: He has 3 cats. P3: He has 3 times as many dogs as cats. P4: He has 2 fewer rabbits than dogs. P5: He also has a collection of gerbils that's 1/3 the number of fish he has. P6: He has a fish tank with three times the number of fish as rabbits.
Narrative Structure P1–narr–P2 P2–narr–P3 P3–narr–P4 P4–narr–P5 P5–narr–P6
</structure>

Answer without DS (generated by GPT-4-Turbo):

1. Larry has 3 cats.
2. He has 3 times as many dogs as cats. 3 * 3 = 9 dogs
3. He has 2 fewer rabbits than dogs. 9 - 2 = 7 rabbits
4. He has 1/3 the number of gerbils as fish. 7 / 3 = 2 gerbils
5. He has 3 times the number of fish as rabbits. 3 * 7 = 21 fish

Total pets: 3 (cats) + 9 (dogs) + 7 (rabbits) + 2 (gerbils) + 21 (fish) = 42 pets

Final Answer: 42.

Answer with DS:

<answer> To determine the total number of pets Larry has, we need to follow the premises in the given order.

From P2, we know Larry has 3 cats.

From P3, we know he has 3 times as many dogs as cats. 3 * 3 = 9 dogs.

From P4, we know he has 2 fewer rabbits than dogs. 9 - 2 = 7 rabbits.

From P6, we know he has a fish tank with three times the number of fish as rabbits. 3 * 7 = 21 fish.

From P5, we know he has a collection of gerbils that's 1/3 the number of fish he has. 1/3 * 21 = 7 gerbils.

Now, we have the number of each type of pet: - Cats: 3 - Dogs: 9 - Rabbits: 7 - Fish: 21 - Gerbils: 7

To find the total number of pets, we add the number of each type of pet: 3 + 9 + 7 + 21 + 7 = 47

Therefore, Larry has 47 pets. </answer>

Question: Marcus is trying to decide whether he really needs to do his homework. If the normal teacher comes in, there's a 40% chance she'll give everyone an extension. There's a 50% chance that tomorrow he'll have a substitute teacher who won't collect the homework. Even if the whole class doesn't get an extension, there's a 20% chance Marcus can convince the teacher his dog ate his assignment and get a personal extension. What is the percentage chance that Marcus will actually have to turn in his homework tomorrow?

Correct Answer: 24%

Structure:

<structure>

Topics [topic-a]: Marcus's homework dilemma [topic-b]: Probabilities of extensions or no extension

Relationships t1–elab–t2

Premises P1: Marcus is trying to decide whether he really needs to do his homework. P2: If the normal teacher comes in, there's a 40% chance she'll give everyone an extension. P3: There's a 50% chance that tomorrow he'll have a substitute teacher who won't collect the homework. P4: Even if the whole class doesn't get an extension, there's a 20% chance Marcus can convince the teacher his dog ate his assignment and get a personal extension.

Narrative Structure P1–narr–P2 P2–narr–P3 P3–narr–P4

</structure>

Answer without DS:

1. The normal teacher comes in and does not give an extension. - 60% chance (since there's a 40% chance she will give one). - Marcus has a 20% chance of getting a personal extension. - 80% chance he will not get the extension and must turn in his homework. - Probability: 60% * 80% = 48%

2. The normal teacher gives an extension (40% chance). - Marcus does not turn in his homework.

3. A substitute teacher comes in (50% chance). - Marcus does not turn in his homework.

4. Total probability of Marcus having to turn in his homework:

- Scenario 1 contributes 48% probability. - Other scenarios do not contribute.

Final Answer: 48% chance that Marcus must turn in his homework.

Answer with DS:

<answer> To find the percentage chance that Marcus will actually have to turn in his homework tomorrow, we need to consider all scenarios where he won't get an extension.

First, we calculate the probability of the normal teacher coming in and not giving an extension: Probability of normal teacher coming in = 100% - 50% (substitute teacher) = 50% Probability of normal teacher not giving an extension = 100% - 40% (giving an extension) = 60% So, the probability of normal teacher coming in and not giving an extension = 50% * 60% = 30%

Now, we calculate the probability of Marcus not getting a personal extension from the normal teacher: Probability of not getting a personal extension = 100% - 20% (getting a personal extension) = 80% So, the probability of normal teacher coming in, not giving an extension, and Marcus not getting a personal extension = 30% * 80% = 24%

Next, we calculate the probability of the substitute teacher coming in and collecting the homework: Probability of substitute teacher coming in = 50% Probability of substitute teacher collecting the homework = 100% (since it's not stated that the substitute teacher won't collect homework in all cases, we assume the worst-case scenario for Marcus)

However, since the problem states that the substitute teacher won't collect the homework, this scenario is actually in Marcus's favor, and we don't need to consider it as a case where he has to turn in his homework.

Now, we add the probability of the normal teacher coming in, not giving an extension, and Marcus not getting a personal extension to get the total probability of Marcus having to turn in his homework: Probability of Marcus having to turn in his homework = 24% (normal teacher, no extension, no personal extension)

Therefore, the percentage chance that Marcus will actually have to turn in his homework tomorrow is 24%. </answer>

Question: A bumper car rink has 12 red cars. They have 2 fewer green cars than they have red cars. The rink also has yellow cars. They have 3 times the number of blue cars as they have green cars. If the rink has 75 cars in total how many yellow cars do they have?

Correct Answer: 23

Structure:

<structure>

Topics [topic-a]: Bumper car rink inventory [topic-b]: Color counts of cars

Relationships t1–elab–t2

Premises P1: A bumper car rink has 12 red cars. P2: They have 2 fewer green cars than they have red cars. P3: The rink also has yellow cars. P4: They have 3 times the number of blue cars as they have green cars. P5: The rink has 75 cars in total.

Narrative Structure P1–narr–P2 P2–narr–P3 P3–narr–P4 P4–narr–P5

</structure>

Answer without DS:

1. There are 12 red cars.

2. There are 2 fewer green cars than red cars. 12 - 2 = 10 green cars

3. The number of yellow cars is 3 times the number of blue cars, but the number of blue cars is not given. Instead, we rely on the total number of cars given.

4. Total number of cars = 75 Non-yellow cars = Red cars + Green cars 12 + 10 = 22 non-yellow cars

5. Yellow cars = Total cars - Non-yellow cars 75 - 22 = 53 yellow cars

Final Answer: 53 yellow cars.

Answer with DS:

<answer>

To find out how many yellow cars the rink has, we need to follow the relationships and premises as structured in the abstract.

From P1, we know the rink has 12 red cars.

From P2, we know they have 2 fewer green cars than they have red cars. 12 - 2 = 10 green cars.

From P4, we know they have 3 times the number of blue cars as they have green cars. 3 * 10 = 30 blue cars.

Now, we know the rink has: - 12 red cars - 10 green cars - 30 blue cars

Total cars accounted for so far: 12 + 10 + 30 = 52 cars.

From P5, we know the rink has 75 cars in total.

Since we've accounted for 52 cars, the remaining cars must be yellow. 75 - 52 = 23 yellow cars.

Therefore, the rink has 23 yellow cars.

</answer>

A.5 GSM-MOD Example

A STORY AND ITS VARIANTS

Original Story

Kenny wants to make sure he does more jumping jacks every week than he did the week before. He recorded that last week he did 324 total jumping jacks. On Saturday of this week, he looks at his records and sees that on Sunday he did 34. On Monday he did 20. On Tuesday he skipped a day. On Wednesday he did 123. On Thursday he did 64. On Friday he did 23. How many does he have to do on Saturday to make sure he beats last week's number?

Variation: C-MOD

Zeta wants to write more lines of code this sprint than she did in the previous sprint. Last sprint, she wrote 324 lines in total. This sprint, so far: Day 1: 34 lines Day 2: 20 lines Day 3: 0 lines Day 4: 123 lines Day 5: 64 lines Day 6: 23 lines How many lines must Zeta write on Day 7 to exceed 324?

Variation: N-MOD

Kenny wants to make sure he does more jumping jacks this week than he did the week before. He recorded that last week he did 648 total jumping jacks. On Saturday of this week, he looks at his records and sees that on Sunday he did 68, on Monday 40, on Tuesday he skipped a day (0), on Wednesday 246, on Thursday 128, and on Friday 46. How many does he have to do on Saturday to ensure he beats last week's number of 648?

Variation: L-MOD

Kenny intends to ensure he performs more star jumps every week than he completed the previous week. He documented that last week he completed 324 overall star jumps. On Saturday of this week, he examines his logs and notices that on Sunday he accomplished 34. On Monday he accomplished 20. On Tuesday he omitted a day. On Wednesday he accomplished 123. On Thursday he performed 64. On Friday he accomplished 23. How many does he have to perform on Saturday to ensure he surpasses last week's total?

Figure 2: Comparison of an original story and its variants.