MAGDI: Structured Distillation of Multi-Agent Interaction Graphs Improves Reasoning in Smaller Language Models

Justin Chih-Yao Chen [*1] **Swarnadeep Saha** [*1] **Elias Stengel-Eskin** [1] **Mohit Bansal** [1]

Abstract

Multi-agent interactions between Large Language Model (LLM) agents have shown major improvements on diverse reasoning tasks. However, these involve long generations from multiple models across several rounds, making them expensive. Moreover, these multi-agent approaches fail to provide a final, single model for efficient inference. To address this, we introduce MAGDI, a new method for *structured distillation of the reasoning interactions between multiple LLMs into smaller LMs*. MAGDI teaches smaller models by representing multi-agent interactions as graphs, augmenting a base student model with a graph encoder, and distilling knowledge using three objective functions: next-token prediction, a contrastive loss between correct and incorrect reasoning, and a graph-based objective to model the interaction structure. Experiments on seven widely-used commonsense and math reasoning benchmarks show that MAGDI improves the reasoning capabilities of smaller models, outperforming several methods that distill from a single teacher and multiple teachers. Moreover, MAGDI also demonstrates an order of magnitude higher efficiency over its teachers. We conduct extensive analyses to show that MAGDI (1) enhances the generalizability to out-of-domain tasks, (2) scales positively with the size and strength of the base student model, and (3) obtains larger improvements (via our multi-teacher training) when applying self-consistency – an inference technique that relies on model diversity.[1]

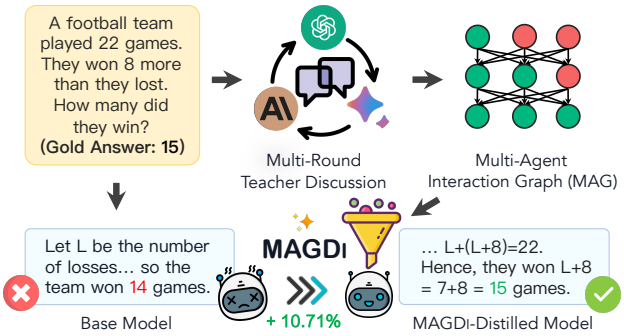

Figure 1. Overview of our distillation method. Given a reasoning problem, multiple teacher-LLMs engage in a multi-round discussion, leading to the generation of a multi-agent interaction graph (MAG). Then our structured distillation method, MAGDI distills reasoning knowledge from these graphs into a base student model.

1. Introduction

Debate and dialogue are natural ways to improve reasoning: we form our best ideas not in isolation, but by refining and discussing them with others. Similarly, we can improve Large Language Models (LLMs) – which often exhibit impressive multi-step reasoning capabilities (Wei et al., 2022; Kojima et al., 2022) – by allowing multiple LLM instances to interact in a discussion (Du et al., 2023; Chen et al., 2023b; Wu et al., 2023). These interactive frameworks enable each agent to iteratively refine its reasoning by obtaining feedback from others, thereby leading to a better consensus at the end of multiple interaction rounds.

Discussion frameworks are typically built on top of proprietary models, e.g., GPT-4, Bard, Claude, etc., which can act as general conversational agents, handle long contexts, and follow instructions (Bubeck et al., 2023). However, these models are also computationally and monetarily expensive, especially when used in multi-round interactions, which require numerous long-token length inference calls to the underlying LLMs.[2] Moreover, these frameworks do not result in a final, joint model that can then be directly used for inference and instead require invoking all inter-

[*]Equal contribution [1]UNC Chapel Hill. Correspondence to: Justin Chih-Yao Chen <cyaochen@cs.unc.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]Code/data: https://github.com/dinobby/MAGDi.

[2]For example, one such multi-LLM interaction method, ReConcile (Chen et al., 2023b) uses around 1900 tokens per sample on a math reasoning task, with other discussion-based methods (e.g., Du et al. (2023); Wu et al. (2023)) using even larger token budgets.

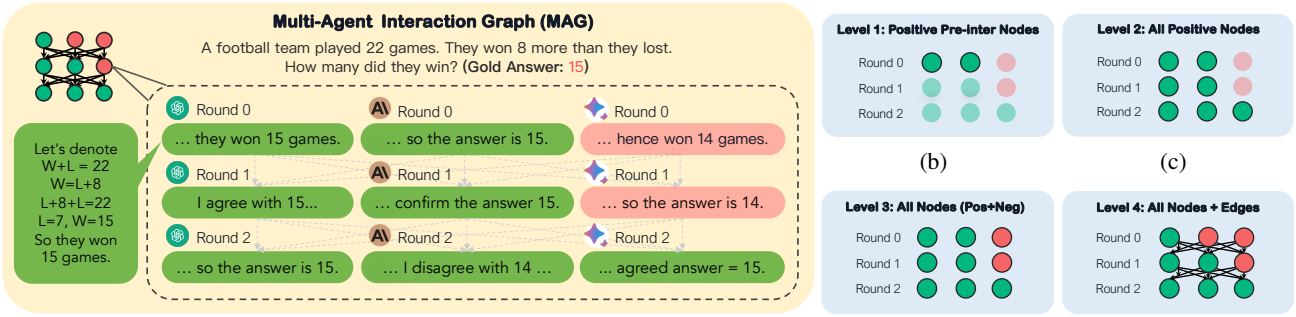

Figure 2. **Left (a):** Illustration of a Multi-Agent Interaction Graph (MAG) constructed with GPT4, Bard, and Claude2 collaboratively solving a math reasoning problem over three discussion rounds. **Right (b-e):** The four levels that characterize our structured distillation method (MAGDI); each level progressively distills knowledge from the highlighted components of a MAG.

acting LLMs at test time. To reduce this cost and train a small, affordable yet capable model, we tackle the problem of teaching reasoning to smaller language models via *structured distillation of the interactions between multiple stronger teacher models*. Specifically, we develop a structured distillation method, **M**ulti-**A**gent Interaction **G**raphs **Di**stillation, or MAGDI, that enables a student model to learn from multi-teacher interactions, with the goal of developing a performant and efficient standalone alternative to expensive multi-agent setups. On seven popular benchmarks in both commonsense and math reasoning, we find increasing improvements over distillation baselines as we incorporate more levels of teacher interactions and structure.

Multi-agent, multi-round interactions are characterized by their participating agents, the number of interaction rounds, and an interaction function defining what information an agent has access to while generating its responses. This function gives rise to a structure between agents and rounds. To learn from this structure, we represent it in **M**ulti-**A**gent Interaction **G**raphs (MAG), a graph-based encoding of multi-agent interactions (§3.1). See Fig. 2(a) for an example. Concretely, a MAG is a directed acyclic graph (DAG) wherein each node represents an agent's generation (in this case, the Chain-of-Thought reasoning (Wei et al., 2022) for a given problem) in a discussion round, annotated with a binary label indicating whether the answer is correct. The edges denote the discussion's structure, indicating which previous turns agents are responding to. MAGs are an intuitive and generalizable way of representing the levels of many multi-agent interactions with varying conversation patterns (Wu et al., 2023), and will allow us to distill this information into a student model for performing zero-shot inference from just the question (i.e., MAGs are not required at test-time).

Given a reasoning problem, MAGs capture rich knowledge of (1) diverse *pre-* and *post-interaction correct reasoning chains* generated by different LLMs (green nodes in Fig. 2(a)), (2) diverse and challenging *incorrect reasoning chains* generated by different LLMs that are refined over interaction rounds (red nodes in Fig. 2(a)), and (3) an *iterative and structured (graph-based) interaction process* that enables this refinement of model reasoning (edges in Fig. 2(a)). We capture all this knowledge via the following four levels of MAG components, which are then used in our distillation method, MAGDI (§3.4), and further tested as part of our experiments (§5.1).

Level 1: Learning from multiple teachers. The student learns from the correct reasoning of *multiple* teachers, rather than one (correct pre-interaction nodes in a MAG, Fig. 2(b)).

Level 2: Learning from teacher interactions. The student learns from both pre- and *post-interaction* data between multiple teachers (all correct nodes in a MAG, Fig. 2(c)).

Level 3: Learning from negative reasoning. The student additionally distills from *negative or incorrect* reasoning from the teacher models (all nodes in a MAG, Fig. 2(d)).

Level 4: Learning from structure. The student learns from the output and *graph-structure* of teacher LLM interactions (all nodes and edges in a MAG, Fig. 2(e)).

Note that each level builds on the prior levels, motivating our main research question:

Research Question: *How can we effectively distill from diverse teacher interactions into a smaller, efficient student model across increasing levels of interaction structure, also demonstrating scalability and generalizability?*

These levels also shape MAGDI, our structured distillation method. MAGDI enables a student model to learn from our graph-structured interaction data (MAGs), with the goal of developing a performant and efficient standalone alternative to expensive multi-agent setups. We first construct a training dataset of MAGs from a high-performing multi-agent discussion framework (Chen et al., 2023b), featuring discussions between three API-based LLMs: GPT-4, Bard, and Claude2 (§3.2). We then develop student models augmented with a Graph Neural Network (GNN) for learning *structure-aware* representations of positive (correct) and

negative (incorrect) reasoning chains and fine-tune them on MAG data. MAGDI's three fine-tuning objectives are aligned to the four levels: (1) next-token prediction (Levels 1-2), (2) a contrastive loss between correct and incorrect reasoning (Level 3), and (3) a graph-based node classification loss (Level 4). These objectives capture all useful signals in MAGs (i.e., teachers' *correct* and *incorrect* reasoning as well as their underlying *conversation structure*). At test time, the distilled model performs zero-shot inference given just the *question* and the *base model* (without the GNN).

We evaluate MAGDI's effectiveness on seven widely-used commonsense (StrategyQA, CommonsenseQA, ARC-c, BoolQ) and math (GSM8K, MATH, SVAMP) reasoning benchmarks, consistently establishing the following findings across datasets and domains:

- **Multi-teacher distillation improves student performance.** When compared directly to distilling from a single teacher, distilling from *multiple teachers* improves performance (Level 1).
- **The value of teacher interactions**: Distilling from the *post-interaction* outputs of teachers further improves students (Level 2).
- **Negative reasoning helps.** Adding a contrastive objective to learn from *incorrect* reasoning provides a valuable signal to the student model (Level 3).
- **Distilling from structure maximizes accuracy.** When MAGDI distills from the first 3 levels *and the structure* of a MAG, the student achieves the highest accuracy, e.g., up to 10% absolute improvement over a zero-shot baseline and up to 4% over the best single-teacher baseline.
- **MAGDI balances performance with efficiency.** MAGDI-distilled models reduce the number of tokens predicted at test time by up to 9x while outperforming all single-teacher distillation baselines.

Building on these results, we further analyze MAGDI along the following axes:

- **Generalizability.** MAGDI can be used to produce a unified joint multi-task learning model that performs well on multiple domains at once and also generalizes well to held-out datasets not seen during training.
- **Scalability.** MAGDI scales positively with the size and strength of the base student model.
- **Diversity.** The output diversity resulting from our multi-teacher training improves self-consistency (Wang et al., 2023), an inference-time ensemble method relying on diverse model answers.

2. Related Work

Knowledge Distillation. Knowledge distillation has proven effective in transferring knowledge from a larger teacher model to a more compact student model (Hinton et al., 2015; Buciluǎ et al., 2006; Chen et al., 2020) including distillation from multiple teacher models (You et al., 2017; Yang et al., 2020). Following recent work, we focus on distillation from *samples* from a model distribution, or symbolic distillation (West et al., 2022), a form of distillation especially common on LLMs, e.g. in instruction tuning (Wang et al., 2023; Taori et al., 2023; Chiang et al., 2023), where instruction-question-answer triples are sourced from a teacher model. In reasoning with LLMs specifically, recent work has distilled reasoning knowledge from a *single* larger teacher model to a smaller student model (Magister et al., 2023; Shridhar et al., 2023; Fu et al., 2023; Ho et al., 2023; Saha et al., 2023; Mukherjee et al., 2023; Mitra et al., 2023; Li et al., 2023; Deng et al., 2023; Liu et al., 2023) using Chain-of-Thought prompting (CoT; Wei et al., 2022) and also, a combination of multiple prompting techniques (Chenglin et al., 2023; Mukherjee et al., 2023; Mitra et al., 2023). Past work has also distilled modular trajectories from a GPT-4 teacher for solving interactive tasks (Chen et al., 2023a; Yin et al., 2023). Overall, different from these single-teacher settings, we delve into the realm of knowledge distillation from *multiple teachers*. Going one step further, we learn from the *interactions* between teachers, bringing in fresh challenges on flexibly representing and modeling these interactions.

Graph-based Interactions. Dialogues, debates, and multi-party conversations (Kirchhoff & Ostendorf, 2003; Leifeld, 2018; Wei et al., 2023) have a long, rich history of being modeled as graphs for different downstream tasks such as emotion and sentiment identification (Ghosal et al., 2019; Shen et al., 2021), dialogue act recognition (Qin et al., 2021), dialogue summarization (Chen & Yang, 2021), and machine reading (Ouyang et al., 2021). Our motivation in this work differs in two major respects. Firstly, we focus on model-model interactions rather than the human-human or human-model interactions from past work. Secondly, our objective is to enhance reasoning capabilities in smaller student models through structured distillation, as opposed to utilizing graph modeling for downstream graph tasks.

3. Method

In Sec. 3.1, we provide a general, formal description of MAG, our graph-based representation of multi-agent interactions. Sec. 3.2 then describes the construction of MAGs for several tasks that will serve as training data for distillation. In Sec. 3.3, we analyze this training data in terms of its structural properties. Lastly, in Sec. 3.4, we describe MAGDI, our structured distillation method for learning from MAGs.

3.1. Multi-Agent Interaction Graph (MAG)

Definition. Consider a collaborative multi-agent setting, where $\mathcal{A} = \{A_i\}_{i=1}^{n}$ agents are interacting with each other

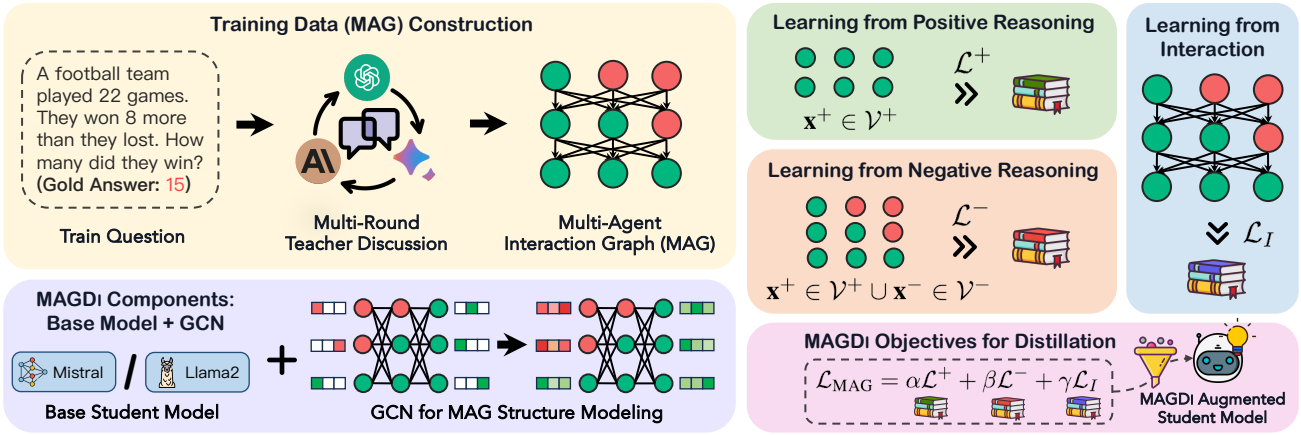

Figure 3. **Training Data Construction:** Given a reasoning problem, multiple teachers go through a multi-round discussion process, generating multi-agent interaction graphs (MAGs). **MAGDI:** Our structured distillation method augments a base student model with a Graph Neural Network (specifically, a GCN) to learn structure-aware representations of reasoning chains. The resultant model is then fine-tuned with a combination of three objectives involving positive chains, negative chains, and the underlying interactions.

for r rounds to solve a task. A multi-agent interaction graph (MAG) is a structured encoding of the interactions between these $\mathcal{A}$ agents over r rounds. Formally, a MAG is a Directed Acyclic Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of directed edges. A node $v_{i,j} \in \mathcal{V}$ represents the output of an agent $A_i \in \mathcal{A}$ in interaction round $j \in [0, r]$.[3] For example, see Fig. 2(a) where each node represents an agent's Chain-of-Thought reasoning to the question. Edges denote conditional dependencies between the agents' interactions and encode how an agent's output is refined over interaction rounds. Specifically, we define a directed edge between two nodes if the target node's generation is conditioned on the source node's generation.

Example of a MAG. MAGs can be generally defined for arbitrary agents and interaction patterns. We focus on reasoning problems as a general class of domains where interaction has positive impacts, defining MAGs for LLM agents on commonsense and math reasoning tasks. Past works have defined several such interaction frameworks (Du et al., 2023; Liang et al., 2023; Chen et al., 2023b) for which MAGs can be defined. Of these, we choose RECONCILE (Chen et al., 2023b) as our interaction framework because of (1) its performance: it obtains the highest performance across multiple benchmarks, and (2) its flexibility: it allows each agent to converse in natural language following the generic Chain-of-Thought reasoning paradigm and thus, can be readily applied to any downstream task where CoT is applicable. MAGs make minimal assumptions about the contents of nodes and hence, can be similarly defined for agents conversing using other prompting techniques (Chen et al., 2022). Fig. 2(a) shows an example of a RECONCILE-based MAG for a math problem. Since the nodes in a MAG represent model responses (either correct or incorrect), we additionally annotate each node $v \in \mathcal{V}$ with a binary label $y_v \in \{0, 1\}$, indicating the correctness of the answer. These are marked with green and red circles in Fig. 2. We refer to these two sets of nodes as $\mathcal{V}^+$ and $\mathcal{V}^-$ respectively such that $\mathcal{V} = \mathcal{V}^+ \bigcup \mathcal{V}^-$. Following the interaction pattern of RECONCILE that conditions each subsequent round of interaction on *all* agent outputs from the previous round, we define edges from all source nodes $v_{i,j}$ to all target nodes $v_{k,j+1}$, i.e., $(v_{i,j}, v_{k,j+1}) \in \mathcal{E}\ \forall i, k \in [1, n], \forall j \in [0, r]$.

[3] Round 0 refers to an agent's pre-interaction output.

3.2. Training Data (MAG) Construction

With the specifics of a MAG defined, we now want to construct these graphs for a given task. These would then serve as data for training distilled models. Given a reasoning problem (e.g., a math word problem), we follow RECONCILE and use GPT-4, Bard, and Claude2 as the three interacting LLM agents for a maximum of three rounds (see Chen et al. (2023b) for further details of the framework). The discussion continues until a consensus is reached, i.e., all agents agree on the same answer. This means that when there is no interaction (i.e., all agents' initial answers are the same), a MAG will have 3 disconnected nodes, one for each agent (and no edges). When there is a single round of interaction, it will have 6 nodes and 9 edges, and so on. In summary, for a given task, our training data will consist of graphs that can be grouped into four structural types, based on the number of interaction rounds. We will refer to these graph types as $\mathcal{G}_0$, $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$ (with the subscript denoting the number of rounds). Fig. 2(a) is an example of a $\mathcal{G}_2$ graph.

Benchmarks. Following the above framework, we construct MAGs for 5 widely-used benchmarks on commonsense and math reasoning: (1) StrategyQA (Geva et al., 2021), (2) CommonsenseQA (Talmor et al., 2019; Aggarwal et al., 2021), (3) AI2 Reasoning Challenge (Clark

et al., 2018), (4) GSM8K (Cobbe et al., 2021) and (5) MATH (Hendrycks et al., 2021). We also experiment with 2 OOD datasets: BoolQ (Clark et al., 2019) and SVAMP (Patel et al., 2021) for which we *do not* construct MAGs and are exclusively used to test the transfer of our model.

3.3. Statistics of Training Data (MAGs)

We construct 1000 training MAGs for each in-domain benchmark. We categorize the data along three dimensions: rounds, agents, and graph structures. See Appendix Table 12 for statistics along each dimension and for each benchmark.

Round. Recall that a MAG consists of nodes belonging to different interaction rounds $i \in [0, 3]$. All datasets have more nodes in lower rounds, indicating that consensus between agents is typically achieved in these earlier rounds. The number of nodes in the later rounds is also an indicator of the difficulty of the benchmark. For example, MATH has the most number of round-3 nodes, suggesting that even after two rounds of discussion, the strong teacher LLMs do not converge on a single answer.

Agent. We can also group the MAG nodes based on the agent generating the response at that node (GPT4/Claude2/Bard). The number of nodes for each agent is the same because all agents engage in all rounds.

Graph Structure. Lastly, we also show the break-down of different MAG structures. Like nodes, $\mathcal{G}_0$ graphs are the most represented in our datasets while $\mathcal{G}_3$ graphs are the least, and all graph structures add up to 1K data points per task. We show examples of MAGs in Appendix D. Using these MAGs as supervision, we train task-specific and multi-task distilled models, as discussed below.

3.4. MAGDI: Structured Distillation from MAGs

We now discuss our proposed structured distillation method, MAGDI, that distills reasoning capabilities into a smaller student model via multi-teacher interaction graphs (MAGs).

MAGDI Overview. Broadly, MAGDI performs structured distillation by augmenting a student model (e.g., Mistral-7B-Instruct or LLaMA-2-7B-Chat) with a light-weight Graph Neural Network (GNN) that is responsible for modeling the 'structure' in structured distillation (see Fig. 3 'MAGDI Components'). This augmented student model is then fine-tuned with a combination of three objectives that distill knowledge from the positive nodes, negative nodes, and the edges (i.e., interactions) in a MAG (see Fig. 3 'MAGDI Objectives for Distillaton'). We denote the base model as $p_\theta(\cdot)$ and the input reasoning problem as q. For brevity, we denote the generation at any MAG node as a variable-length sequence of tokens: $\mathbf{x} = \{x_1, x_2, ..., x_{|\mathbf{x}|}\}$. Below, we describe the three objectives for structurally distilling student LMs.

Objective 1: Learning from Positive Reasoning. Learning from a teacher LLM's *correct* reasoning has been shown to improve smaller models (Magister et al., 2023; Li et al., 2023). Hence, MAGDI first fine-tunes the student model on all *correct* reasoning chains $\mathbf{x}^+ \in \mathcal{V}^+$ using a standard next-token prediction objective (see Fig. 3 'Learning from Positive Reasoning'). The loss is defined as follows.

$$\mathcal{L}^+ = -\sum_{\mathbf{x}^+ \in \mathcal{V}^+} \sum_{i=1}^{|\mathbf{x}^+|} \log p_\theta(x_i^+ | \mathbf{x}_{<i}^+, q) \tag{1}$$

Objective 2: Learning from Negative Reasoning. Teacher LLMs also make mistakes, particularly when solving more challenging problems. However, instead of discarding these, MAGDI treats them as *challenging* negatives (generated by a strong teacher) that a student model can learn from by contrasting with positive chains. Given a positive reasoning chain $\mathbf{x}^+ \in \mathcal{V}^+$ and a negative reasoning chain $\mathbf{x}^- \in \mathcal{V}^-$, we first extract representations of these chains by performing a weighted average pool over the final layer's hidden representations of the constituent tokens. We represent these as $h_{\mathbf{x}^+} \in \mathbb{R}^d$ and $h_{\mathbf{x}^-} \in \mathbb{R}^d$ respectively where d is the embedding dimension. Using a projection matrix and a $\tanh$ activation, we further project these embeddings to two scalar scores $s_{\mathbf{x}^+} \in [-1, 1]$ and $s_{\mathbf{x}^-} \in [-1, 1]$. MAGDI then optimizes for the following margin-based objective (Cortes & Vapnik, 1995) for pairs of positive and negative chains $\{\mathbf{x}^+, \mathbf{x}^-\} \in \mathcal{V}^+ \times \mathcal{V}^-$ in a MAG (see Fig. 3 'Learning from Negative Reasoning').

$$\mathcal{L}^- = \sum_{\mathbf{x}^+ \in \mathcal{V}^+} \sum_{\mathbf{x}^- \in \mathcal{V}^-} \max(0, \rho - s_{\mathbf{x}^+} + s_{\mathbf{x}^-}) \tag{2}$$

where $\rho \in [-1, 1]$ is the margin (set to 1 in our experiments).

Objective 3: Learning from Interaction. Beyond distillation from the correct and incorrect nodes, MAGDI is also intended to distill from the entire conversational *structure* present in the teachers' discussion. This would allow the student model to summarize the discussion process and acquire knowledge of how a teacher refines its reasoning chain in each discussion round. Hence, while the previous two objectives assumed nodes to be a disconnected set, MAGDI removes this assumption by also modeling the edges. MAGDI achieves this by augmenting the base student model with a Graph Convolution Network (GCN) module (Kipf & Welling, 2017). The goal of the GCN is to learn improved, 'structure-aware' representations of reasoning chains (nodes) such that the student model learns to discriminate between correct and incorrect nodes in a MAG, eventually leading to better generation of reasoning chains.

Given any positive or negative node $\mathbf{x} \in \mathcal{V}$, MAGDI generates a node representation $h_{\mathbf{x}} \in \mathbb{R}^d$ from the base LM. It then learns structure-aware representations of these nodes

with a two-layer GCN using the following equation:

$$h_{\mathbf{x}}^{(l+1)} = \sigma(D^{-1}Mh_{\mathbf{x}}^{(l)}W^{(l)})$$

where $M \in |\mathcal{V}| \times |\mathcal{V}|$ is the adjacency matrix with self-connections, D is the diagonal degree matrix of M, σ is the ReLU activation function, $h_{\mathbf{x}}^{(l)}$ and $h_{\mathbf{x}}^{(l+1)}$ are the input and updated node representations of the l-th layer. We set $h_{\mathbf{x}}^{(0)} = h_{\mathbf{x}}$ and $W^{(l)}$ as the weight matrix of the l-th layer. After two layers of message passing, we obtain the final node representation $h_{\mathbf{x}}^{(L)}$, which is now conditioned on the graph structure. MAGDI then projects $h_{\mathbf{x}}^{(L)}$ with a linear layer parameterized by $W_c \in \mathbb{R}^{d \times C}$ (where C is the number of node labels) and applies the softmax function to derive the probability distribution over the node labels $\hat{y}_{\mathbf{x}} = \text{softmax}(h_{\mathbf{x}}^{(L)}W_c)$. Finally, we use cross-entropy loss for the (correct/incorrect) node classification objective over all nodes in a MAG (see Fig. 3 'Learning from Interaction'),

$$\mathcal{L}_I = -\sum_{\mathbf{x}\in\mathcal{V}}\sum_{i=1}^{C} y_{\mathbf{x}}^{(i)} \log(\hat{y}_{\mathbf{x}}^{(i)}) \tag{3}$$

where $y_{\mathbf{x}}$ is a one-hot encoding of the label of a node $\mathbf{x}$.

Final Objective. Our final loss, $\mathcal{L}_{MAG}$, as defined below, is a weighted combination of the three losses.

$$\mathcal{L}_{MAG} = \alpha\mathcal{L}^+ + \beta\mathcal{L}^- + \gamma\mathcal{L}_I \tag{4}$$

with $\alpha, \beta, \gamma \in [0, 1]$ being the respective weights.

MAGDI Inference. The GCN module is only used during the distillation process. Hence, at test time, the student model performs zero-shot inference given just the question and uses only the same number of parameters and architecture as the base student model.

4. Experimental Setup

Implementation Details. We test our structured distillation method, MAGDI, with three instruction-tuned student models that are of different scales and belong to different model families: Mistral-7B-Instruct, LLaMA-2-7B-Chat, and LLaMA-2-13B-Chat. We train both task-specific distilled models (i.e., trained and tested on a single downstream task) and also one joint multi-task distilled model (trained on all in-domain tasks together and then tested on each in-domain and out-of-domain task). The multi-task model represents our unified model for OOD tasks. For conciseness, we refer MAGDI to the *resultant task-specific models* and MAGDI-MT will refer to the *resultant multi-task model*. See Appendix A for other implementation details.

Baselines. We group all methods into three categories of 'Distillation Source Type', described as follows. **(1) No Teacher.** The lower bound of our distilled models is the zero-shot base model (e.g., Mistral-7B-Instruct). **(2) Single-Teacher.** Our next set of baselines only uses training data from a single teacher (out of the three agents used to construct MAGs). This follows multiple prior works (Li et al., 2023; Magister et al., 2023; Fu et al., 2023; Ho et al., 2023) that fine-tune student models on CoT reasoning using the next-token prediction objective (Equation 1). In particular, we fine-tune three student models with one of GPT-4, Bard, or Claude2 as the teacher using *only* the positive samples from the respective teacher model. We will refer to these **Si**ngle-**T**eacher distilled models as SIT. Then SIT-GPT4, for example, will refer to a distilled model trained with only GPT4 as the teacher. **(3) Multi-Teacher.** Due to the lack of existing multi-teacher baselines, we first adapt an existing single-teacher method, Distilling Step-by-Step (Hsieh et al., 2023) to a multi-teacher setup. The other multi-teacher baselines (MAGDI-*) correspond to distilled models trained with increasing levels of MAGs; these baselines demonstrate the utility of the levels as defined in Fig. 2.

- **DSS-MT.** Hsieh et al. (2023) propose 'Distilling Step-by-Step (DSS)' with a multi-task objective to predict the label and rationale separately. We apply the same multi-task objective to all teachers to directly compare the effectiveness of the DSS objectives on the same MAG data. We refer to this as DSS-MT (DSS with Multi-Teacher).
- **MAGDI-R0 (Level 1).** MAGDI-R0 refers to a model that is fine-tuned only on the *Round-0* (pre-interaction) correct reasoning of multiple teachers (i.e., GPT-4, Bard, and Claude2) using only the $\mathcal{L}^+$ objective defined in Eqn. 1, i.e., Level 1 of MAGDI.
- **MAGDI-CN (Level 2).** Next, MAGDI-CN is a distilled model that is trained on *all correct* (pre- and post-interaction) reasoning from all teachers with again the same $\mathcal{L}^+$ loss, corresponding to Level 2 of MAGDI.
- **MAGDI-AN (Level 3).** Going one step further, MAGDI-AN is a distilled model that is trained on *all nodes* (correct and incorrect) from all teachers. Hence, this is trained with both $\mathcal{L}^+$ and $\mathcal{L}^-$ objectives in Equations 1 and 2. Note that all three models for Levels 1-3 are *unstructured* multi-teacher distilled models that view MAGs as a set of correct and incorrect nodes.
- **MAGDI (Level 4).** This is our final full method that distills knowledge from *all nodes and edges* of a MAG using all three objectives as described in Equation 4.

5. Results and Analysis

5.1. Main Results

Our primary results demonstrate the effectiveness of MAGDI across five reasoning benchmarks over different

Table 1. Comparison of structured distillation (MAGDI) with no teacher, single-teacher, and multi-teacher distillation baselines. Firstly, MAGDI outperforms *all* baselines across *all* five reasoning benchmarks. On average, MAGDI outperforms the strongest SIT-GPT4 baseline by 4.61% and the no teacher baseline by 10.71%. Secondly, knowledge distillation from each component of MAG improves the student model, as demonstrated by a consistent increase in performance from Level 1 to Level 4. Lastly, we also make all nodes available for a multi-teacher baseline, DSS-MT. MAGDI obtains a larger improvement over SIT-GPT4 than DSS-MT does (2.49% vs. 4.61%).

Distillation Type	Distillation Data	Distilled Model	Datasets					Average Acc
			StrategyQA	CSQA	ARC-c	GSM8K	MATH	
No Teacher (Jiang et al., 2023)	-	Mistral-7B-Instruct	61.57	57.89	60.32	44.05	7.02	46.17
Single-Teacher (Li et al., 2023; Magister et al., 2023; Fu et al., 2023; Ho et al., 2023)	Claude2	SIT-Claude2	64.39	64.18	68.24	45.34	7.24	49.89
	Bard	SIT-Bard	68.56	65.06	66.87	45.61	7.06	50.63
	GPT-4	SIT-GPT4	69.96	66.87	68.91	47.38	8.24	52.27
Multi-Teacher	All Nodes	DSS-MT (Hsieh et al., 2023)	71.18	69.42	71.38	51.84	9.98	54.76 [+ 2.49%]
	Round-0 Nodes	MAGDI-R0 [Level 1]	71.18	67.36	72.06	48.52	9.72	53.77 [+ 1.50%]
	Correct Nodes	MAGDI-CN [Level 2]	71.62	69.31	72.34	50.11	10.66	54.81 [+ 2.54%]
	All Nodes	MAGDI-AN [Level 3]	72.10	70.65	71.92	50.69	11.98	55.47 [+ 3.20%]
	MAG	MAGDI [Level 4]	**74.24**	**72.56**	**72.61**	**52.27**	**12.76**	**56.88 [+ 4.61%]**

single-teacher and multi-teacher distillation setups. For the main results, we use Mistral-7B-Instruct as the student model and train task-specific distilled models (see Sec. 5.2 for experiments with multi-task models with larger students belonging to different model families). We report accuracy for each task. Based on Table 1 results, we summarize our main conclusions below, addressing our research question posed in Section 1: How can we distill diverse teacher interactions into a smaller and more efficient student, utilizing the *increasing levels of interaction structure in a MAG*?

Level 1: Distillation from multiple teachers outperforms distillation from the single strongest teacher. Knowledge distillation from the correct reasoning chains of multiple teachers i.e., GPT-4, Bard, and Claude2, outperforms distillation from the single strongest teacher by an average of 1.50% (see MAGDI-R0 row versus SIT-GPT4 row). Different teachers bring diversity in their reasoning, leading to improved reasoning capabilities of the student model.

Level 2: Distillation from pre- and post-interaction reasoning outperforms only pre-interaction reasoning. Training a student model on all correct reasoning chains after multiple teacher models have interacted further improves the student's reasoning. Fine-tuning on post-interaction reasoning chains effectively increases the training data and its usefulness is validated through an overall 2.54% improvement, compared to the best single-teacher model (see MAGDI-CN row versus SIT-GPT4 row).

Level 3: Negative reasoning chains help. Additionally, learning by contrasting between positive and negative chains improves over all previous levels. Our MAGDI-AN model obtains an overall average improvement of 3.20% (see MAGDI-AN row versus SIT-GPT4 row).

Level 4: Structured distillation from interactions outperforms all multi-teacher baselines. Our final structured distillation method, MAGDI, that distills knowledge from both the outputs (nodes) and structure (edges) of multi-agent interactions, obtains the largest improvement and outperforms all single- and multi-teacher models. When applied to Mistral-7B-Instruct, MAGDI surpasses a model that learns only from GPT4 by an average of 4.61%. Out of all the levels, distillation from interactions brings the most improvements (both individually and across all five benchmarks), specifically demonstrating the utility of modeling the conversation structure of multi-agent discussions. Overall, compared to the base model, MAGDI improves the student's reasoning capabilities by a significant 10.71% (46.17 $\rightarrow$ 56.88) while maintaining similar inference-time efficiency.

MAGDI also outperforms adapted single-teacher distillation from prior work. When using data from all teacher models, MAGDI consistently outperforms DSS-MT. Since the learning source is the same in both methods (i.e., all nodes in MAGs), it underscores the effectiveness of learning from teacher interactions and our proposed objectives.

MAGDI improves inference efficiency compared to RECONCILE. Compared to RECONCILE, MAGDI also drastically improves inference efficiency. Because the sizes of gated models like those used in RECONCILE are not known, we measure efficiency via number of output tokens generated.[4] As shown in Table 2, MAGDI achieves up to a 9x reduction in token count. MAGDI obtains this efficiency by always answering questions in one inference call; this differs from RECONCILE which has to go through up to twelve (expensive) LLM inference calls (involving 3 agents for the initial round and another 3 rounds). Moreover, Figure 4 and Table 10 in Appendix shows the trade-off between token efficiency and performance, averaged across datasets and for each individual dataset respectively. While RECONCILE (as the upper-bound) has the best performance, it

[4]This metric is extremely strict and under-estimates our efficiency gains, as each of the gated LLMs used by RECONCILE exceeds 7B parameters, and each example for RECONCILE involves multiple inference calls.

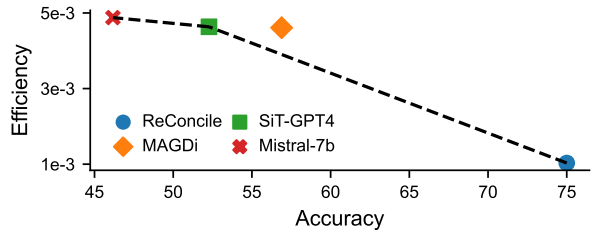

Figure 4. Trade-off between performance and efficiency. MAGDI exceeds the Pareto frontier of prior work, surpassing single-teacher models in performance and surpassing RECONCILE in efficiency, defined as $1/avg(tokens)$.

Table 2. Comparison of the token counts generated by RECONCILE (a multi-agent interaction framework) and MAGDI.

	RECONCILE	MAGDI	Reduction
StrategyQA	924.5	107.5	8.6x
CSQA	936.9	104.2	9.0x
ARC-c	448.3	86.4	5.2x
GSM8K	642.3	141.6	4.5x
MATH	1900.1	645.0	2.9x
Average	970.4	216.9	4.5x

also produces the most tokens (cf. Table 2); the zero-shot and single-teacher models are more efficient than RECONCILE but suffer in terms of performance. On the other hand, MAGDI achieves a better balance of efficiency and performance than these baselines, with more efficiency than RECONCILE and higher performance than the zero-shot and prior single-teacher distillation methods.

MAGDI effectively transfers multi-agent capabilities into a single student. In Table 3, we compare distilled student models to their teachers by reporting results for (1) GPT4, as the upper-bound performance of a single-teacher, (2) RECONCILE, as the upper-bound of multi-agent teacher (using GPT4, Claude2, Bard), (3) SiT-GPT4, as distillation from the single-teacher, and (4) MAGDI as distillation from multi-agent teacher. We also report the relative improvement from 'single' to 'multi', both with and without distillation. The relative improvement from SiT-GPT4 to MAGDI (12.70%) is much higher than that from GPT-4 to the multi-teacher system RECONCILE (5.79%), highlighting the effectiveness of MAGDI in transferring multi-agent capabilities into a single student model.

5.2. Analysis: Generalizability, Scalability, Diversity

MAGDI can be used to train one joint multi-task model. In our main experiments (Sec. 5.1), we fine-tuned task-specific models with structured distillation. While these multi-teacher task-specific models show clear benefits over single-teacher models, we would ideally like *one joint* model that can tackle all tasks. Therefore, we train a joint multi-task model (MAGDI-MT) by combining training data from all five benchmarks and evaluating it together on all

Table 3. Comparison of single-teacher and multi-teacher distilled students with their respective teachers. The relative improvement from SiT-GPT4 to MAGDI (12.70%) is higher than that from GPT-4 to the multi-teacher system ReConcile (5.79%).

	StrategyQA	CSQA	ARC-c	GSM8K	MATH	Avg
GPT4	75.60	73.30	94.50	90.70	39.00	74.62
ReConcile	87.70	78.70	96.30	92.10	41.00	79.16
% improved	13.80	6.86	1.87	1.52	4.88	**5.79**
SiT-GPT4	69.96	66.87	68.91	47.38	8.24	52.27
MAGDI	74.24	72.56	72.61	52.27	12.76	56.88
% improved	5.77	7.84	5.10	9.36	35.42	**12.70**

Table 4. Out-of-domain comparison between Single-Teacher Multi-Task (SIT-GPT4-MT) and MAGDI Multi-Task (MAGDI-MT) models. MAGDI-MT performs up to 7% better than the single-teacher baseline even on OOD datasets (57.52 vs. 64.30).

	BoolQ	SVAMP
SiT-GPT4-MT	60.70	57.52
MAGDI-MT	**63.98**	**64.30**

benchmarks. MAGDI-MT obtains an average accuracy that is within 1% of task-specific models (56.89% versus 55.12%), showing its applicability for training a joint model (refer to Appendix Table 13 for the full table).

MAGDI generalizes to OOD tasks. We also evaluate MAGDI-MT on two out-of-domain benchmarks (BoolQ for commonsense reasoning and SVAMP for math) that were not included in multi-task training. As shown in Table 4, MAGDI outperforms single-teacher distillation by to 3% on BoolQ and 7% on SVAMP. The broader implication of this is that the single takeaway MAGDI-MT model maintains good performance on OOD tasks and continues to outperform single-teacher baselines on new datasets.

MAGDI scales positively with better base models. We now study the scaling properties of structured distillation by varying the base student model. In particular, we train three structurally distilled models with MAGDI, using LLaMA-2-7B-Chat, LLaMA-2-13B-Chat, and Mistral-7B-Instruct as the base models. In Fig. 5, we plot the average accuracy of all five benchmarks (and see Table 7 for individual results). The ordering of the three models is based on their zero-shot performance. We observe that MAGDI brings consistent improvements over all base models, maintaining proportionate gains. Overall, the scaling trend of MAGDI suggests that structured distillation should continue to improve even stronger students. See Appendix B for other analyses of MAGDI e.g., the effect of the amount of training data and different graph structures.

MAGDI boosts self-consistency. We hypothesize that a student model that learns from *multiple teachers* will exhibit better diversity in its generations. To test this, we combine MAGDI with an ensemble method like Self-Consistency (SC) (Wang et al., 2023). SC computes a majority vote

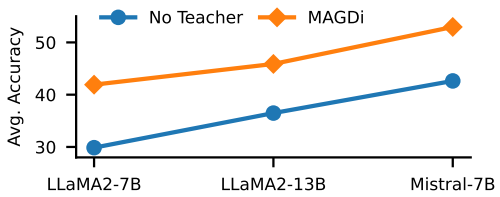

Figure 5. Scaling results of MAGDI with different base student models. As the average (zero-shot) performance of the base model improves (Mistral-7B > LLaMA-2-13B> LLaMA-2-7B), MAGDI shows a corresponding increase.

over model answers and has proven effective for reasoning tasks; SC's improvements are predicated on answer diversity (since a majority vote over the versions of the same answer would not yield any improvements). We show that self-consistency with our *multi-teacher distilled model* outperforms the same with (1) the base model and more importantly, (2) a single-teacher distilled model. As shown in Table 5, on GSM8K, SC between 10 responses improves accuracy by 15% when applied to a MAGDI-trained student, compared to only 4% improvement for the base model (Mistral-7B-Instruct) and 11% for the single-teacher model (SIT-GPT4). Broadly, this suggests that inference-time algorithms relying on a model's inherent diversity can be boosted from multi-teacher distillation.

GCN outperforms self-attention for modeling MAGs. To demonstrate the effectiveness of the GCN in encoding MAGs, we compare it to two alternative approaches. These remove the explicitly-defined graph structure of the GCN (which is determined by the MAG) and instead aim to automatically learn the interactions with multi-head attention layers (Vaswani et al., 2017). *Token-level attention:* Here we linearize the entire interaction (across agents and rounds) into a single sequence of tokens, so that any token (across all reasoning chains, rounds, and agents) can attend to any other token. *Node-level attention:* Here we keep the nodes intact, i.e., each node still represents a whole reasoning chain, but we linearize them into a sequence of nodes so that any node can attend to any other node. Results presented in Table 6 show that visualizing the entire interaction as a sequence of tokens results in long token lengths and removes the boundaries between the nodes (reasoning chains), leading to significantly worse performance (first row). Node-level attention improves slightly over token-level attention by keeping the nodes (chains) intact and modeling the interactions between them with self-attention layers (second row). However, our GCN-based approach performs the best across all datasets. While the transformer could theoretically model any graph structure, it may also require a large amount of training data and tuning to infer the correct interaction graph structure. Hence, we find that defining the interaction patterns with MAGs apriori and modeling them with a GCN is more performant and data-efficient. Note that MAGDI's main contributions are in defining structured representations of multi-agent interactions and then distilling such interaction data into a weaker student model. This contribution is independent of the exact graph representation learning module, and going forward, we hope that our results will motivate newer methods of modeling these multi-agent interactions.

Table 5. Self-consistency with MAGDI on GSM8K achieves the largest gain (up to 15%) compared to the same with the base student model and the single-teacher distilled model.

	Mistral-7B	SIT-GPT4	MAGDI
w/o SC	44.05	47.38	52.27
w/ SC	48.44 [+ 4.39%]	58.62 [+ 11.24%]	**67.42 [+ 15.15%]**

Table 6. Comparison of different graph modeling methods. MAGDI with GCN consistently outperforms attention-based variants for modeling interaction graphs.

Method	StrategyQA	CSQA	ARC-c	GSM8K	MATH
Token Attn	68.56	65.04	70.31	51.17	8.86
Node Attn	69.87	71.16	70.01	51.63	9.22
GCN	**74.24**	**72.56**	**72.61**	**52.27**	**12.76**

6. Discussion and Conclusion

We have tackled the problem of structured distillation from multi-agent interactions as a way to equip much smaller and more efficient language models with improved reasoning capabilities. To achieve this goal, we proposed a graph-based representation of these interactions, generated graphs for training, and developed a structured distillation method for learning from these interaction graphs. Our results showed the effectiveness, generalizability, and scalability of structured distillation across multiple reasoning benchmarks.

While MAGDI relies on interactions between LLMs for supervision, its modular design makes it well-suited even to scenarios where such data may be limited. Revisiting MAGDI's four objectives, we split its use cases into the following four categories. (1) *No teacher data is available:* When no teacher data is available, we demonstrate MAGDI's ability to generalize in a zero-shot manner (Table 4). (2) *Only single-teacher data is available:* When supervision from only one teacher model is available, we can leverage the first objective in MAGDI (Level 1). (3) *Having multi-teacher data but no interaction:* When multiple teacher sources are present but the interactions are absent, we show gains from the first two MAGDI objectives (Level 3). (4) *Having multi-teacher interaction data:* This is when the full MAGDI model (Level 4) can be applied to maximize performance. Thus, our four levels succinctly depict the adaptability of MAGDI based on the amount and nature of data availability, providing a structured framework for implementing its objectives across varying data scenarios and applications.

Acknowledgements

We thank Peter Hase and Archiki Prasad for useful feedback and suggestions regarding experiments. This work was supported by NSF-CAREER Award 1846185, NSF-AI Engage Institute DRL-2112635, DARPA MCS Grant N66001-19-2-4031, Accelerate Foundation Models Research program, and a Google PhD Fellowship. The views contained in this article are those of the authors and not of the funding agency.

Impact Statement

The computing resources used by LLMs incur a substantial carbon footprint (Strubell et al., 2019), and running them in interactive multi-agent settings like RECONCILE further increases these environmental costs. MAGDI distills from several LLMs into a single more efficient and smaller LM, thus reducing the environmental cost of running LLMs while still maintaining strong performance.

The LLMs MAGDI distills from – and the student models it distills into – can reflect stereotypes, biases, and other negative traits present in their pre-training data (Weidinger et al., 2021), which we do not have control over. Our distilled LLMs have the same capacity for undesirable generation as their teacher models and their respective zero-shot variants; as such, the models resulting from MAGDI distillation have the same potential for misuse as any LLM or method distilling from LLMs. More studies are needed to evaluate and mitigate such biases in LLMs.

References

Aggarwal, S., Mandowara, D., Agrawal, V., Khandelwal, D., Singla, P., and Garg, D. Explanations for CommonsenseQA: New Dataset and Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3050–3065. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.acl-long.238/.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. URL https://arxiv.org/abs/2303.12712.

Buciluǎ, C., Caruana, R., and Niculescu-Mizil, A. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 535–541, 2006. URL https://dl.acm.org/doi/abs/10.1145/1150402.1150464.

Chen, B., Shu, C., Shareghi, E., Collier, N., Narasimhan, K., and Yao, S. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*, 2023a. URL https://arxiv.org/abs/2310.05915.

Chen, J. and Yang, D. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1380–1391, 2021. URL https://aclanthology.org/2021.naacl-main.109/.

Chen, J. C.-Y., Saha, S., and Bansal, M. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*, 2023b. URL https://arxiv.org/abs/2309.13007.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. URL https://arxiv.org/abs/2006.10029.

Chen, W., Ma, X., Wang, X., and Cohen, W. W. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*, 2022. URL https://arxiv.org/abs/2211.12588.

Chenglin, L., Qianglong, C., Caiyu, W., and Yin, Z. Mixed distillation helps smaller language model better reasoning. *arXiv preprint arXiv:2312.10730*, 2023. URL https://arxiv.org/abs/2312.10730.

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1300. URL https://aclanthology.org/N19-1300.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20:273–297, 1995.

Deng, Y., Prasad, K., Fernandez, R., Smolensky, P., Chaudhary, V., and Shieber, S. Implicit chain of thought reasoning via knowledge distillation. *arXiv preprint arXiv:2311.01460*, 2023. URL https://arxiv.org/abs/2311.01460.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate, 2023. URL https://arxiv.org/abs/2305.14325.

Fu, Y., Peng, H., Ou, L., Sabharwal, A., and Khot, T. Specializing smaller language models towards multi-step reasoning. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10421–10430. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/fu23d.html.

Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., and Berant, J. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 154–164, 2019.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021. URL https://arxiv.org/abs/2103.03874.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Ho, N., Schmid, L., and Yun, S.-Y. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.830. URL https://aclanthology.org/2023.acl-long.830.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-k., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, July 2023.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Kirchhoff, K. and Ostendorf, M. Directions for multi-party human-computer interaction research. In *Proceedings of the HLT-NAACL 2003 Workshop on Research Directions in Dialogue Processing*, pp. 7–9, 2003. URL https://aclanthology.org/W03-0703/.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022. URL https://arxiv.org/abs/2205.11916.

Leifeld, P. Discourse network analysis. policy debates as dynamic networks,[w:] jn victor, ah montgomery, m. lubell (red.), 2018.

Li, L. H., Hessel, J., Yu, Y., Ren, X., Chang, K.-W., and Choi, Y. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2665–2679, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.150. URL https://aclanthology.org/2023.acl-long.150.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., and Shi, S. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023. URL https://arxiv.org/abs/2305.19118.

Liu, B., Bubeck, S., Eldan, R., Kulkarni, J., Li, Y., Nguyen, A., Ward, R., and Zhang, Y. Tinygsm: achieving >80% on gsm8k with small language models. *arXiv preprint arXiv:2312.09241*, 2023. URL https://arxiv.org/abs/2312.09241.

Magister, L. C., Mallinson, J., Adamek, J., Malmi, E., and Severyn, A. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1773–1781, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.151. URL https://aclanthology.org/2023.acl-short.151.

Mitra, A., Del Corro, L., Mahajan, S., Codas, A., Simoes, C., Agarwal, S., Chen, X., Razdaibiedina, A., Jones, E., Aggarwal, K., et al. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*, 2023. URL https://arxiv.org/abs/2311.11045.

Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., and Awadallah, A. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*, 2023. URL https://arxiv.org/abs/2306.02707.

Ouyang, S., Zhang, Z., and Zhao, H. Dialogue graph modeling for conversational machine reading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3158–3169, 2021. URL https://arxiv.org/abs/2012.14827.

Patel, A., Bhattamishra, S., and Goyal, N. Are NLP models really able to solve simple math word problems? In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168.

Qin, L., Li, Z., Che, W., Ni, M., and Liu, T. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 13709–13717, 2021. URL https://arxiv.org/abs/2012.13260.

Saha, S., Hase, P., and Bansal, M. Can language models teach weaker agents? teacher explanations improve students via personalization. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2306.09299.

Shen, W., Wu, S., Yang, Y., and Quan, X. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1551–1560, 2021. URL https://arxiv.org/abs/2105.12907.

Shridhar, K., Stolfo, A., and Sachan, M. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7059–7073, 2023. URL https://arxiv.org/abs/2212.00193.

Strubell, E., Ganesh, A., and McCallum, A. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3645–3650, 2019.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/

3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language models with self-generated instructions. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837, 2022. URL https://arxiv.org/abs/2201.11903.

Wei, J., Shuster, K., Szlam, A., Weston, J., Urbanek, J., and Komeili, M. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*, 2023. URL https://arxiv.org/abs/2304.13835.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

West, P., Bhagavatula, C., Hessel, J., Hwang, J., Jiang, L., Le Bras, R., Lu, X., Welleck, S., and Choi, Y. Symbolic knowledge distillation: from general language models to commonsense models. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4602–4625, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.341. URL https://aclanthology.org/2022.naacl-main.341.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Zhang, S., Zhu, E., Li, B., Jiang, L., Zhang, X., and Wang, C. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023. URL https://arxiv.org/abs/2308.08155.

Yang, Z., Shou, L., Gong, M., Lin, W., and Jiang, D. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 690–698, 2020. URL https://arxiv.org/abs/1910.08381.

Yin, D., Brahman, F., Ravichander, A., Chandu, K., Chang, K.-W., Choi, Y., and Lin, B. Y. Lumos: Learning agents with unified data, modular design, and open-source llms. *arXiv preprint arXiv:2311.05657*, 2023. URL https://arxiv.org/abs/2311.05657.

You, S., Xu, C., Xu, C., and Tao, D. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1285–1294, 2017. URL https://dl.acm.org/doi/pdf/10.1145/3097983.3098135.

Appendix

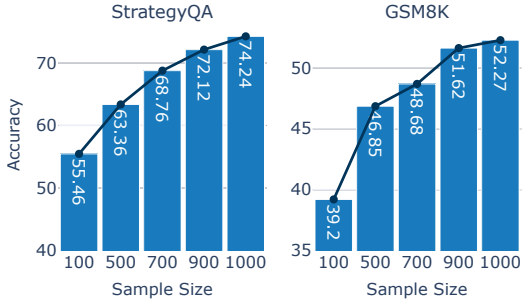

Figure 6. Results with scaling training data on StrategyQA and GSM8K. With structured distillation, student accuracy increases with an increase in training data.

A. Implementation Details of MAGDI

Learning from Negative Reasoning. We provide additional details about MAGDI's contrastive objective in Equation 2. In practice, since a MAG can have many nodes, computing the contrastive loss for every possible positive and negative pair can be prohibitively expensive. In such cases, we randomly sample nodes from the minority group to pair up with the nodes from the majority group. For instance, if there are 5 positive and 3 negative nodes in a MAG, then it naturally forms three pairs and for the remaining 2 positives, we randomly sample 2 negatives to pair with them.

Model Training. MAGDI-distilled models are fine-tuned using Low-Rank Adaptation (LoRA) for efficiency (Hu et al., 2022). We set the rank to 16 and alpha to 32. We fine-tune the student model for 10 epochs using a learning rate of 5e-6 and a batch size of 16. For the hyperparameters α, β and γ, please refer to our code implementation for detailed settings in each dataset. All of our experiments are run on four RTX A6000 with 48G memory each.

Table 7. Results of MAGDI with different base student models (LLaMA-2-7B-Chat, LLaMA-2-13B-Chat, and Mistral-7B-Instruct). First, across all benchmarks and for all three base models, MAGDI outperforms no teacher baselines by a large margin. Second, the effect of structured distillation correlates with the performance of the base model (Mistral-7B > LLaMA-2-13B > LLaMA-2-7B), highlighting the scaling properties of MAGDI.

Base	Distilled	StrategyQA	CSQA	GSM8K	MATH
LLaMA-2-7B	No Teacher	51.53	46.81	18.60	2.50
	MAGDi	**66.81**	**66.73**	**28.36**	**5.76**
LLaMA-2-13B	No Teacher	58.52	51.73	31.77	3.90
	MAGDi	**69.00**	**67.05**	**40.56**	**6.46**
Mistral-7B	No Teacher	61.57	57.89	44.05	7.02
	MAGDi	**74.24**	**72.56**	**52.27**	**12.76**

Table 8. Dataset licenses

Dataset	License
StrategyQA	MIT License (License)
CommonsenseQA	MIT License (License)
ARC-c	CC BY-SA 4.0 (License)
GSM8K	MIT License (License)
MATH	MIT License (License)
BooQ	CC BY-SA 3.0 (License)
SVAMP	MIT License (License)

B. Further Analyses of MAGDI

MAGDI scales positively with better base models. In Table 7, we show MAGDI's scaling trends on StrategyQA, CommonsenseQA, GSM8K and MATH. Here we apply MAGDI to three base models: Mistral-7B-Instruct, LLaMA-2-7B-Chat, and LLaMA-2-13B-Chat. Across all these benchmarks, MAGDI demonstrates consistent gains on top of all base models.

MAGDI scales positively with the amount of training data. Next, we analyze the scaling properties of MAGDI by varying the amount of training data. We train distilled models with MAGDI (using Mistral-7B-Instruct as the base model) by varying the amount of training data from 100 to 1000 samples. Figure 6 shows that with more data, MAGDI exhibits better performance – e.g., on StrategyQA, training on 1K MAGs improves reasoning performance by 10.88% compared to training on 500 samples. This suggests that MAGDI may bring additional improvements with a larger training corpus.

Dense interaction graphs distill significant knowledge to students. Recall that the MAGs in our corpus have distinct structures, with $\mathcal{G}_3$ being the densest graph (having the most nodes and edges). We show that removing these $\mathcal{G}_3$ graphs from our training corpus leads to a significant drop in student accuracy. As shown in Table 9, structured distillation on CSQA without the $\mathcal{G}_3$ graphs causes student performance to drop by 2% and additionally removing the $\mathcal{G}_2$ graphs causes a further drop of 1%. Our result thus highlights the importance of distillation from denser graphs (even if they are sparsely represented in the corpus).

MAG edges effectively model the discussion structure. Recall that we defined directed edges in MAGs according to the interaction pattern between agents across rounds. To further demonstrate the utility of these edges, we compare our directed MAGs (D-MAG) to fully-connected MAGs (FC-MAG) where every node is connected to every other node – and undirected MAGs (UD-MAG) where the edges follow the interaction but are undirected. As shown in Table 11, FC-MAG performs significantly worse, showing the utility of defining edges according to interaction patterns. Finally, the directionality of the edges has marginal impact on the final performance.

Table 9. Removing denser interaction graphs (e.g., $\mathcal{G}_3$ or $\mathcal{G}_2$) from CSQA training corpus leads to a significant drop in student accuracy, highlighting the importance of learning from such graphs.

Training MAGs	Accuracy
All	72.56
w/o $\mathcal{G}_3$	70.65
w/o $\mathcal{G}_2$ & $\mathcal{G}_3$	69.65

Table 10. Dataset-wise breakdown of the performance and efficiency trade-off shown in Fig 4.

	StrategyQA	CSQA	ARC-c	GSM8K	MATH	Avg
# Tok (ReConcile)	924.5	936.9	448.3	642.3	1900.1	970.4
# Tok (MAGDi)	107.5	104.2	86.4	141.6	645.0	216.9
Reduction	8.6x	9.0x	5.2x	4.5x	2.9x	4.5x
Acc (ReConcile)	79.0	74.7	93.5	85.3	41.0	74.7
Acc (MAGDi)	74.2	72.6	72.6	52.3	12.8	56.9

Table 11. Utility of defining edges according to interaction structure in MAGs. Fully-connected MAGs perform worse than defining (directed or undirected) edges based on interactions.

Method	StrategyQA	CSQA	ARC-c	GSM8K	MATH
FC-MAG	71.59	71.22	69.38	50.29	11.02
UD-MAG	74.02	**72.61**	72.46	52.10	12.66
D-MAG	**74.24**	72.56	**72.61**	**52.27**	**12.76**

Table 12. Training and test statistics for five benchmarks. Our training corpus (i.e., MAGs) is categorized along three dimensions: (1) **Round:** number of nodes belonging to each interaction round, (2) **Agent:** number of nodes belonging to each agent (GPT4/Bard/Claude2), and (3) **Graph:** number of graphs with a specific graph structure.

Task	Train			Test
	Round (0 / 1 / 2 / 3)	Agent (Each)	Graph ($\mathcal{G}_0$ / $\mathcal{G}_1$ / $\mathcal{G}_2$ / $\mathcal{G}_3$ / All)	
StrategyQA	3K / 843 / 438 / 102	1.4K	719 / 135 / 112 / 34 / **1K**	229
CSQA	3K / 2.1K / 942 / 627	2.2K	306 / 380 / 105 / 209 / **1K**	2.2K
ARC-c	3K / 792 / 288 / 153	1.4K	736 / 168 / 45 / 51 / **1K**	1.1K
GSM8K	3K / 1.3K / 588 / 354	1.7K	557 / 247 / 78 / 118 / **1K**	1.3K
MATH	3K / 2.3K / 1.5K / 1.2K	2.7K	215 / 269 / 89 / 427 / **1K**	5K

Table 13. Comparison between task-specific MAGDI models and the joint multi-task MAGDI model (MAGDI-MT). On average, MAGDI-MT's accuracy is within 1% of task-specific models.

Distilled Model	StrategyQA	CSQA	ARC-c	GSM8K	MATH	Average Accuracy
MAGDI-MT	70.74	71.73	72.87	49.51	10.76	55.12
MAGDI	74.24	72.56	72.61	52.27	12.76	56.89

C. Benchmark Licenses

The licenses of the seven datasets we used are in Table 8.

D. Qualitative Examples of MAG

See Fig. 7 and Fig. 8 for two examples of MAGs.

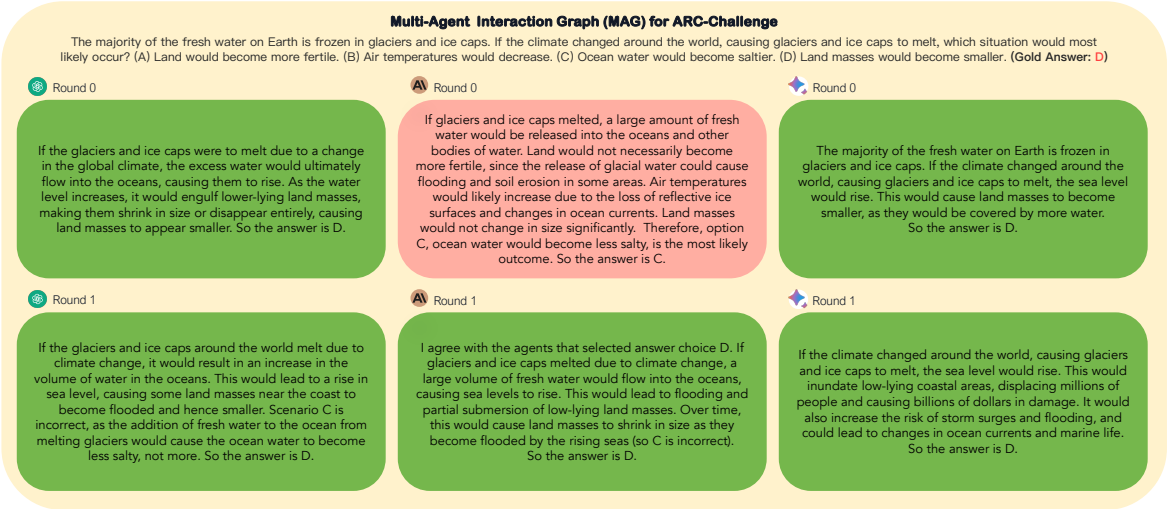

Figure 7. An example of a MAG from the ARC-c dataset.

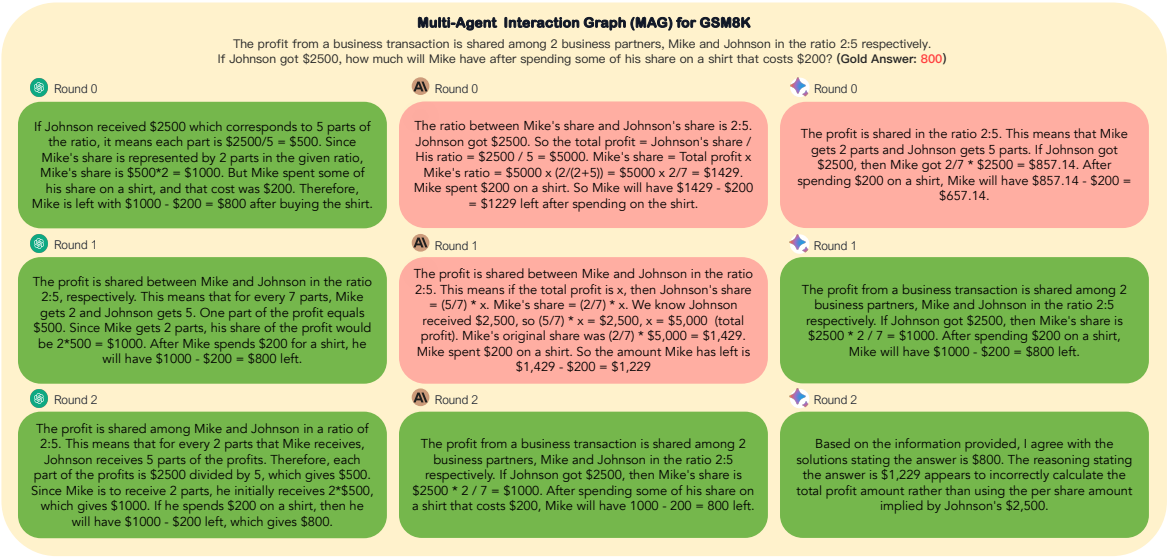

Figure 8. An example of a MAG from the GSM8K dataset.