SMAGDi: Socratic Multi Agent Interaction Graph Distillation for Efficient High Accuracy Reasoning

Aayush Aluru* aayush.aluru09@gmail.com
Myra Malik* maliknmyra@gmail.com
Samarth Patankar* samarth.patankar10@gmail.com
Spencer Kim
Kevin Zhu kevin@algoverseairesearch.org
Sean O'Brien
Vasu Sharma

Algoverse AI Research

Abstract

Multi-agent systems (MAS) often achieve higher reasoning accuracy than single models, but their reliance on repeated debates across agents makes them computationally expensive. We introduce *SMAGDi*, a distillation framework that transfers the debate dynamics of a five-agent Llama-based MAS into a compact Socratic decomposer-solver student. SMAGDi represents debate traces as directed interaction graphs, where nodes encode intermediate reasoning steps with correctness labels and edges capture continuity and cross-agent influence. The student is trained with a composite objective combining language modeling, graph-based supervision, contrastive reasoning, and embedding alignment to preserve both fluency and structured reasoning. On StrategyQA and MMLU, SMAGDi compresses a 40B multi-agent system into a 6B student while retaining *88% of its accuracy*, substantially outperforming prior distillation methods such as MAGDi, standard KD, and fine-tuned baselines. These results highlight that explicitly modeling interaction graphs and Socratic decomposition enable small models to inherit the accuracy benefits of multi-agent debate while remaining efficient enough for real-world deployment.

1 Introduction

Large Language Models (LLMs) have proven adept at generation, problem-solving, translation, and summarization, performing well on various benchmarks. An LLM's capability to learn from vast amounts of data allows it to handle complex tasks logically and fluently. Although LLMs have demonstrated competency in handling diverse tasks, their performance is often hurt by inaccuracies (Zhang et al., 2023).

Multi-agent systems (MAS) have been explored as a solution to combat this issue. These systems leverage collaborative reasoning among multiple agents to enhance accuracy through diverse perspectives (Ramirez-Medina et al., 2025; Bersenev et al., 2024; Talebirad and Nadiri, 2023).

However, MASs introduce substantial computational overhead due to repeated interactions across multiple rounds of debate. (Zhang et al., 2024). This creates scalability barriers for deployment in traditional environments, as they lack a unified model for efficient inferences since traditional distillation methods fail to capture the nuances of reasoning processes.

*Lead Authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

MAGDi (Multi-Agent Graph Distillation) is an advancement addressing these limitations by employing graph-based representations to capture multi-agent reasoning traces (Chen et al., 2024). It utilizes three objective functions to distill and demonstrate gains over single-teacher methods.

Despite this, MAGDi suffers from limitations that constrain its holistic effectiveness. It inadequately replicates the remarkably diverse reasoning mechanics of a MAS because it is a single-agent system (SAS), leading to a reduction in accuracy (Curran et al., 2023; Peng et al., 2024). This hinders its ability to preserve the nuances of collaboration, leading to an incomplete knowledge transfer from the teacher system.

Our work proposes SMAGDi (Socratic Multi-Agent Graph Distillation), a novel framework that integrates Socratic Chain of Thought (SCoT) distillation with multi-agent graph representations to overcome MAGDi's limitations. Unlike standard knowledge distillation methods focusing primarily on output replication, Socratic CoT emphasizes cross-agent collaboration and systematic problem decomposition (Shridhar et al., 2023). SCoT utilizes a modular dual-architecture consisting of a problem decomposer and subproblem solver, enabling more effective reasoning transfer compared to monolithic approaches like MAGDi (Kang et al., 2024; Jacobs and Jordan, 1990).

Our approach addresses the **Research Question**: *How can a model distilled through Socratic Chain of Thought from a multi-agent system teacher optimize computational efficiency while maintaining superior reasoning capabilities?*

SMAGDi enhances MAGDi by integrating SCoT to capture the back-and-forth reasoning dynamics of multi-agent systems. While MAGDi falls short in distilling debate depth and collaborative reasoning patterns, SMAGDi leverages SCoT's problem decomposition capabilities to better mirror the interactive validation processes that make MAS effective at reducing inaccuracies.

The framework employs five specialized agents for the MAS with distinct personas (Lawyer, Scientist, Mathematician, Ethicist, and Historian) to ensure comprehensive domain coverage, while utilizing dynamic weighting mechanisms that prioritize domain-relevant expertise.

We present our benchmark statistics, methodology, evaluation process, and key findings, providing insights into weightage-based multi-agent settings and model distillation in applied ML settings.

Contributions introduced:

- Persona-based, dynamically weighted MAS for cross-domain performance
- SCoT distillation technique, with a four-term loss component, for transferring debate processes of a MAS into low-parameter, step-by-step reasoning models

2 Related Works

2.1 Knowledge Distillation (KD)

Knowledge distillation is a machine learning technique in which a large and complex "teacher" model transfers its knowledge to a smaller "student" model. Standard Knowledge Distillation (SKD) uses a two-part cross-entropy loss on hard and soft labels. The hard labels represent ground truths, while soft labels are the teacher's logits. This forces the model to mimic the probability distributions of the teacher model's logits, allowing the student to obtain similar next-token predictions as the teacher. This effectively distills both accurate and in-depth reasoning into a smaller model. (Hinton et al., 2015).

Instead of always transferring the reasoning processes, standard KD models are susceptible to merely mimicking the responses, making the models infeasible in situations where advanced reasoning capabilities are required.

2.2 Socratic Chain-of-Thought Distillation (SCoT)

Socratic Chain-of-Thought (SCoT) is a knowledge distillation technique that trains its student model on rationals elicited from the teacher model when solving a particular problem, training the student to simplify and solve complex problems in a similar step-by-step manner to the teacher. Unlike other forms of KD, SCoT employs two student models working in tandem as a single unit to simplify

and solve problems using step-by-step reasoning. In this system, one model (a "decomposer") is responsible for breaking down a complex query into several smaller parts, resembling a *Socratic* style of thinking, while its partner (a "solver") solves each problem and pieces together a final output. This type of distillation works best on substantially smaller LLMs and has a higher accuracy rate compared to competitors, specifically over reasoning datasets (Shridhar et al., 2023).

However, what distillation methods like SCoT lack is targeted and optimized usage. SCoT distillation used independently only uses one teacher's logic, which is subject to its ability to answer effectively.

2.3 MAGDi

Multi-agent methods have long been known to provide better responses on complex tasks compared to single-agent systems, but lack the efficiency needed for them to be feasible. MAGDi proposes the first solution to this by distilling an existing multi-agent system into a smaller model. The system creates directed graphs (Multi-Agent Graphs, or MAGs), where nodes and edges capture the agent's reasoning chains. A graph convolutional network is used to distill the processes into a smaller model, which then performs zero-shot inference. (Chen et al., 2024). However, MAGDi uses a single-agent system, which has been proven to perform worse than modular systems in reasoning situations(Curran et al., 2023; Peng et al., 2024).

3 Methodology

In this work, we developed SMAGDi (Socratic Multi-Agent Interaction Graph Distillation), enabling efficient deliberation over multi-step problems by simulating domain-specific personas through SCoT (Socratic Chain-of-Thought) to mimic these analysis processes. These domain-specific personas differ from the traditional use of interchangeable agents, as each has a specific alignment and purpose (Wei et al., 2023).

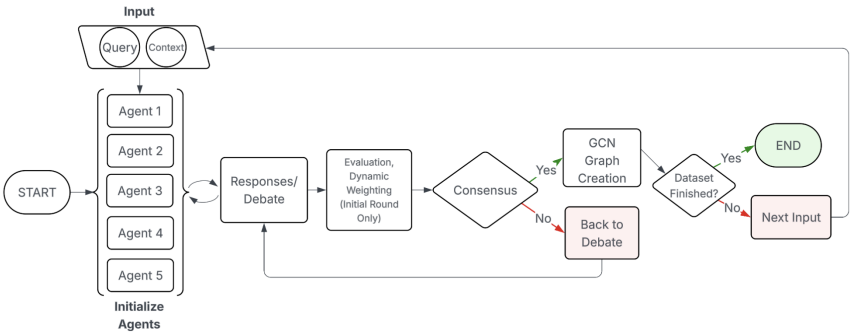

Figure 1: The overarching training pipeline for the creation of SMAGDi's Multi-Agent Interaction Graphs (MAGs) with dynamic weighting, graph construction, and consensus mechanisms.

3.1 Multi-Agent System Setup

Given any natural language problem P_i, SMAGDi aims to predict the most accurate response by amalgamating reasoning and responses from a diverse group of agents. Each agent (A_j) has a different persona - Lawyer, Scientist, Mathematician, Ethicist, and Historian - with role-aligned prompting. The agents will respond, respectively, to the same input (P_i) each analyzing the prompt through role alignment. As seen in Figure 1, once consensus is reached, the responses received from each agent during the debate will be logged onto an interaction graph mapping relationships and progressions of responses.

3.1.1 Dynamic Weight Optimization

After agents are assigned their domain-specific instructions, we implement an initial performance-based agentic weighting mechanism that dynamically adjusts the influence based on each agent's

performance in training examples before graph creation. The weight optimization is calculated as follows:

$$w_i = \max(\varepsilon, \text{accuracy}_i)$$

$$\tilde{w}_i = \frac{w_i}{\sum_j w_j}$$

Where $\tilde{w}_i$ represents the normalized weights, the summation of weights equals one, and $\varepsilon = 0.1$ represents the minimum amount of influence an agent can have during the debate. On the contrary, this scheme permits agents that perform well to exert higher influence over the final output, allowing for optimized performance. This process runs once before the agentic debate.

3.1.2 Debate Process

The deliberation process implements a layered consensus algorithm with dynamic temperature scaling, increasing in later rounds (i.e., +0.1 for every round). In the initial round, agents respond to a question, similar to the process in weight optimization. However, they also provide an analysis of their responses, as context for further debate in later rounds. In these last rounds, models are asked to refine their responses based on the other agents, where more influence is given to agents with higher weights, allowing for the development of strong and complex reasoning paths until consensus is reached or until three rounds are completed, where then, weighted voting is done (Sandwar et al., 2025; Chen et al., 2024).

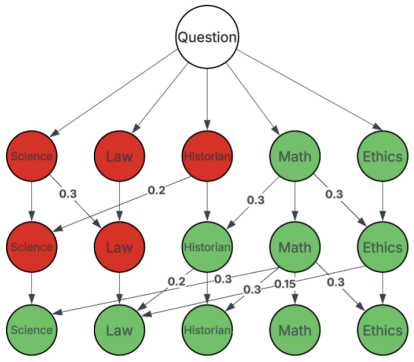

Figure 2: NetworkX graph representation with weighted influence edges, and node correctness for GCN traversal during distillation

3.1.3 Graph Construction

Debate graph creation with NetworkX processes the multi-agent interactions into structured data representations, through a node and edge construction algorithm, as seen in Figure 2. The graphs are established with foundational nodes representing the question. Agents' nodes are annotated with edges connecting them to the foundational nodes. In the ensuing debate, two types of relationships are encoded with continuity edges, connecting consecutive responses from the same agent, and influence edges, capturing cross-agent reasoning influences in a directed format (Chen et al., 2024). These influence edges are captured with the initial trained weights from dynamic weighting, creating a weighted influence network that incorporates the agent's credibility, allowing for the creation of strong reasoning chains. The NetworkX representations are converted to PyTorch Geometric graphs for GCN processing. (Kipf and Welling, 2017). Semantic embeddings capture textual context with the all-mpnet-base-v2 sentence transformer model, providing dense vector representations of agent responses and reasoning. Geometric embeddings are employed concurrently to preserve topological relations via Laplacian positional encoding to ensure the structural context of debates remains intact (Maskey et al., 2022). Ground truth annotation establishes node-level correctness labels enabling supervised classification objectives. The pipeline implements padding with attention masks to accommodate variable graph sizes in multi-agent debates, ensuring effective data.

3.2 SMAGDi

SMAGDi (Socratic Multi-Agent Interaction Graph-Distillation) is a distillation framework that involves the transfer of interactions between multiple LLMs to a smaller model. The respective edges denote a logical path between these interactions, and a graphical representation described above allows the student model to distill this information easily. The application of SMAGDi provides insight into the way each agent (A_j)'s interactions affect the student model, not only increasing efficiency intrinsically but also the system's accuracy. By distilling knowledge into a student model, this framework enables the development of efficient and computationally conservative systems.

3.2.1 Distillation

Once the PyTorch Geometric graphs are constructed from the multi-agent debate process, graph structure and examples consisting of positive and negative reasoning chains and decomposer/solver-specific responses (described below) are passed to the student model for training (Chen et al., 2024; Shridhar et al., 2023). Graph structure $G = (V, E)$ encodes the dependencies between the agent's reasoning steps, where each node $v_i \in V$ represents a reasoning state with features x_i derived from response embeddings, and edges $e_{ij} \in E$ capture influence relationships with weights w_{ij} reflecting agent credibility scores.

The distillation operates through four synergistic loss components derived from four different example types in the graphs that capture different aspects of reasoning: positive (correct reasoning chains), negative (incorrect reasoning chains), examples of (decomposer) sub-reasoning questions, and (solver) reasoning responses.

The Socratic MAGDi model is trained using a composite objective that supervises both linguistic fluency and structured reasoning quality. The total loss combines four components: (1) language modeling for generative fluency, (2) graph-based classification for structural correctness, (3) contrastive loss for distinguishing valid from invalid reasoning, and (4) alignment loss for embedding consistency across model components. Three of these loss components are derived from MAGDi - language modeling (MAGDi's next-token prediction), graph-based node classification, and contrastive reasoning (Chen et al., 2024). We added our alignment loss to ensure agreement between our decomposer and solver models.

Language Modeling Loss (α-weighted)

The language modeling component ensures that both the decomposer and solver produce syntactically and semantically coherent outputs. It uses the standard causal (auto-regressive) cross-entropy loss, encouraging the model to predict each token in a sequence given only the preceding tokens.

$$L_{\text{LM}} = \frac{1}{T} \sum_{t=1}^{T} -\log P(x_t \mid x_{<t}; \theta)$$

Where:

- x_t is the ground truth token at position t,
- $x_{<t}$ is the sequence of tokens before t,
- T is the length of the sequence,
- θ denotes the model parameters.

Node Classification Loss (β-weighted)

To explicitly supervise the structural correctness of reasoning steps, we use a graph-based node classification loss. A GCN processes nodes representing intermediate reasoning steps. The loss is formulated as a binary cross-entropy over the graph nodes:

$$L_{\text{node}} = -\sum_{i=1}^{|V|} \Big[y_i \log(\sigma(h_i^{(L)})) + (1 - y_i) \log(1 - \sigma(h_i^{(L)})) \Big]$$

Where:

- $|V|$ is the number of nodes in the graph,

- $y_i \in \{0, 1\}$ is the ground truth correctness label for the node i,
- $h_i^{(L)}$ is the final embedding of the node i after L GCN layers,
- $\sigma(\cdot)$ is the sigmoid activation function that produces a probability estimate.

This loss ensures that the model generates fluent reasoning and learns to recognize whether individual reasoning steps are factually and logically correct within the graph structure.

Contrastive Reasoning Loss (γ-weighted)

We apply a margin-based ranking loss to reinforce the model's ability to distinguish high-quality reasoning from poor or incorrect alternatives. This loss operates over positive and negative reasoning pairs, encouraging the model to score correct reasoning chains higher than incorrect ones.

$$L_{\text{contrast}} = \frac{1}{N} \sum_{i=1}^{N} \max(0, 1 - s_i^+ + s_i^-)$$

Where:

- s_i^+ and s_i^- are scalar scores assigned to positive (valid) and negative (invalid) reasoning chains, respectively,
- N is the number of paired examples.

The loss penalizes the model when the negative reasoning is scored too similarly or higher than the positive reasoning, enforcing a margin of 1.

Alignment Loss (δ-weighted)

Even if the decomposer and solver perform well independently, their internal representations must remain semantically aligned. To enforce consistency between their reasoning processes, we introduce an embedding alignment loss based on mean squared error (MSE) between their hidden states.

$$L_{\text{align}} = \frac{1}{N} \sum_{i=1}^{N} \left\| z_i^{\text{dec}} - z_i^{\text{sol}} \right\|_2^2$$

Where:

- z_i^{dec} and z_i^{sol} are the projected embedding vectors of the decomposer and solver for example i,
- N is the batch size.

Summary

These four weights are used in conjunction to train the student model, with the total loss being modeled as:

$$L_{\text{total}} = \alpha L_{\text{LM}} + \beta L_{\text{node}} + \gamma L_{\text{contrast}} + \delta L_{\text{align}}$$

Here, $\alpha, \beta, \gamma, \delta$ are hyperparameters that control the contribution of each loss to the total. The loss coefficients found to work best were 1.0, 1.0, 0.1, and 0.5, respectively. It is important to note that these were not tuned extensively.

3.2.2 SCoT

After training, the decomposer and solver agents will recursively generate and answer sub-questions built off P_i, through zero-shot inferencing. This process allows the model to mimic multi-hop reasoning without external supervision after initialization.

4 Experimental Setup

4.1 Benchmarks and Models

We use the following two datasets to test the performance of our and our competitors' models: **(1)** the **StrategyQA** benchmark of general analytical accuracy (Geva et al., 2021), **(2)** and the **MMLU** benchmark of general knowledge (Hendrycks et al., 2021).

To establish a standardized performance metric, we chose to use models purely from the Llama 3 collection. Agents in the MAS are the Llama 3.1-8B-Instruct model, bringing pretrained knowledge through Llama's Instruct finetuning method. The decomposer and solver in the student unit are both Llama 3.2-3B models, and to test the scaling abilities of our system, we performed a brief ablation using the Llama 3.2-1B model (Grattafiori et al., 2024).

4.2 Baselines

To evaluate the performance of SMAGDi, we ran three other models and systems as baselines. We tested MAGDi, Standard KD (SKD), and a finetuned baseline (F-Baseline) of the Llama 3.2-3B model using an identical setup. Our F-Baseline was finetuned using LoRA (Low-Rank Adaptation) from the PEFT (Parameter-Efficient Finetuning) library (Hu et al., 2021). While our SKD was done by aggregating the logits of the MAS.

All models are evaluated by exact match accuracy, as both datasets are either boolean or multiple choice. Across both datasets, models were evaluated using the same data. For StrategyQA, models trained on 80% of the data and tested on the remaining 20%; for MMLU, models trained on 1000 examples from the auxiliary train split and tested on the 14000 testing split. Additionally, all models were trained with a learning rate of 5e-5 and tested on the same number of epochs.

For SMAGDi and MAGDI, to ensure that our distillation process was represented in the purest form, we used zero-shot inference on the test data. This reinforces the system's effectiveness, as models were ensured to learn only from the MAG architecture rather than being influenced by the dataset.

5 Results and Analysis

5.1 MAS and SAS

Before testing our distillation process, we needed to establish that our MAS performed better than a SAS. We tested both the StrategyQA and MMLU datasets on the Llama 3.1-8B-Instruct. For the MAS, we used zero-shot prompting, while the SAS used LoRA fine-tuning. Our results for the testing data are shown in Table 1.

	StrategyQA	MMLU
SAS	77.9	69.4
MAS	**84.2**	**76.4**

Table 1: Baseline and multi-agent system results with Llama 3.1-8B-Instruct showing MAS outperforms SAS by an average of 7%.

We found that the use of a MAS provides higher accuracy than the use of an SAS, +6.3% on StrategyQA and +7% on MMLU. Notably, other experiments comparing multi and single-agent systems show similar results, including a 38% higher accuracy in a MAS relative to a SAS (88% vs 50%) in an ultimatum game, with an insignificant difference in runtime costs (Sreedhar and Chilton, 2024; Gao et al., 2025).

Based on our primary results, we were able to answer our research question presented in Section 1: *How can a model distilled through Socratic Chain of Thought (modular system) from multi-agent systems optimize computational efficiency while maintaining superior reasoning capabilities?* Our results illustrate the effectiveness of Socratic MAGDi as an efficient and high-performing system across two benchmarks, against various distillation methods.

5.2 Distillations

Table 2 and its visual representation (Graph 3) show that SMAGDi outperformed MAGDi, Standard Knowledge Distillation (SKD), and our fine-tuned baseline on benchmark tests. This supports our hypothesis that when multi-agent interaction graphs are distilled socratically, accuracy is boosted, even when compared to a model fine-tuned for the dataset. Additionally, when comparing the models' parameters to the results from Table 1, SMAGDi retained 88% of the MAS's accuracy while reducing parameters by almost 7 times, distilling a 40B system into a 6B decomposer-solver unit.

	StrategyQA	MMLU
3B F-Baseline	67.9	62.4
3B Standard KD	62.7	63.0
3B MAGDi	68.3	64.0
3B **SMAGDi**	**74.6**	**66.1**
1B F-Basline	59.4	49.0
1B **SMAGDi**	**62.4**	**52.2**

Table 2: Performance metrics for different strategies (percentages rounded to the nearest tenth), showing SMAGDi outperforms all existing baselines across both datasets.

Method 1: Fine-Tuned Baseline Compared to the fine-tuned baseline model, our SMAGDI model improved accuracy on both the StrategyQA and MMLU benchmark. Our SMAGDi model had a +6.7% for Strategy QA and +3.7% for MMLU. This resulted in a 5.2% average increase in accuracy for SMAGDi compared to the baseline. These accuracy differences are notable as the baselines are fine-tuned for these datasets.

Method 2: Standard Knowledge Distillation When tested against SKD, our model had a 12.9% accuracy increase in the Strategy QA benchmark, and a 3.1% increase in the MMLU benchmark. Although Socratic Distillation is a newer concept, its dual-model approach and Socratic-style questioning outperform standard cross-entropy loss comparisons.

Method 3: Multi-Agent Interaction Graph Distillation Compared to MAGDi, SMAGDi had a 7.3% increase in accuracy on the StrategyQA dataset. Similarly, there was a 2.1% increase in accuracy when tested in the MMLU benchmark. Our Socratic Distillation structure was able to improve the original MAGDi pipeline by enhancing the model's reasoning capabilities.

1B Model A model must be applied to different scenarios to be considered adequate. As seen in Table 2, our pipeline scales well over differently sized models too, +3.0% for StrategyQA and +3.2% for MMLU. As the 1B model has fewer parameters, the ability of our pipeline to effectively transfer knowledge highlights its effectiveness in low-compute environments.

Summary : The averages of these statistics proved that SMAGDi provides the optimal way to distill an MAS into a student while improving accuracy. SMAGDi not only outperformed MAGDi and SKD but also outperformed our fine-tuned baseline model. This not only shows that our distillation method is better than existing methods, but also shows its ability to capture reasoning patterns effectively.

6 Conclusion

We found that encoding debate traces as structured reasoning signals using Socratic graph distillation provides an effective solution for transferring these signals into a smaller, more efficient model. Through experiments on both the StrategyQA and MMLU datasets, we find that this model can generalize effectively, enabling the deployment of reasoning-capable small models in resource-constrained settings. Broadening the possibilities of distillation allows models to learn more from teachers than ever before, paving the way for future explorations of SOTA-distillation methods.

7 Limitations

Training Data Due to computational constraints, we did not train our models on the full auxiliary train split of MMLU. Further testing will be needed to confirm that SMAGDi will perform better than other models if trained on all data.

Large-Scale Setups Since SMAGDi was tested on lightweight models, we lack proof of its scalability to larger environments. In application, alternatives that provide similar results with minimal computational costs would prove to be more efficient.

Dynamic Weighting and Personas Our MAS used dynamic-weighting mechanisms and persona-based agents as a baseline to distill knowledge into the student. However, MAGDi used a different MAS for their paper without these mechanisms, meaning that we can not ascertain that MAGDi's results were comprehensive and reflected the system's full capabilities.

References

Dennis Bersenev, Ayako Yachie-Kinoshita, and Sucheendra K. Palaniappan. Replicating a high-impact scientific publication using systems of large language models, 2024. URL https://www.biorxiv.org/content/early/2024/04/12/2024.04.08.588614.1.

Justin Chih-Yao Chen, Swarnadeep Saha, Elias Stengel-Eskin, and Mohit Bansal. Magdi: Structured distillation of multi-agent interaction graphs improves reasoning in smaller language models, 2024. URL https://arxiv.org/abs/2402.01620.

Shawn Curran, Sam Lansley, and Oliver Bethell. Hallucination is the last thing you need, 2023. URL https://arxiv.org/abs/2306.11520.

Mingyan Gao, Yanzi Li, Banruo Liu, Yifan Yu, Phillip Wang, Ching-Yu Lin, and Fan Lai. Single-agent or multi-agent systems? why not both?, 2025. URL https://arxiv.org/abs/2505.18286.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies, 2021. URL https://arxiv.org/abs/2101.02235.

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. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Robert A. Jacobs and Michael I. Jordan. A competitive modular connectionist architecture. In *Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3*, NIPS-3, page 767–773, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN 1558601848.

Junmo Kang, Leonid Karlinsky, Hongyin Luo, Zhen Wang, Jacob Hansen, James Glass, David Cox, Rameswar Panda, Rogerio Feris, and Alan Ritter. Self-moe: Towards compositional large language models with self-specialized experts, 2024. URL https://arxiv.org/abs/2406.12034.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017. URL https://arxiv.org/abs/1609.02907.

Sohir Maskey, Ali Parviz, Maximilian Thiessen, Hannes Stärk, Ylli Sadikaj, and Haggai Maron. Generalized laplacian positional encoding for graph representation learning, 2022. URL https://arxiv.org/abs/2210.15956.

Binghui Peng, Srini Narayanan, and Christos Papadimitriou. On limitations of the transformer architecture, 2024. URL https://arxiv.org/abs/2402.08164.

Joaquin Ramirez-Medina, Mohammadmehdi Ataei, and Alidad Amirfazli. Accelerating scientific research through a multi-llm framework. In *Accelerating Scientific Research Through a Multi-LLM Framework*, 2025. URL https://api.semanticscholar.org/CorpusID:276287299.

Vivaan Sandwar, Bhav Jain, Rishan Thangaraj, Ishaan Garg, Michael Lam, and Kevin Zhu. Town hall debate prompting: Enhancing logical reasoning in llms through multi-persona interaction, 2025. URL https://arxiv.org/abs/2502.15725.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. Distilling reasoning capabilities into smaller language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.441. URL https://aclanthology.org/2023.findings-acl.441/.

Karthik Sreedhar and Lydia Chilton. Simulating human strategic behavior: Comparing single and multi-agent llms, 2024. URL https://arxiv.org/abs/2402.08189.

Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents, 2023. URL https://arxiv.org/abs/2306.03314.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. Multi-party chat: Conversational agents in group settings with humans and models, 2023. URL https://arxiv.org/abs/2304.13835.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for llm-based multi-agent systems, 2024. URL https://arxiv.org/abs/2410.02506.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models, 2023. URL https://arxiv.org/abs/2309.01219.

A Testing Set Up

For the SMAGDi distillation, the Hugging Face trainer enabled early stopping with check-pointing to ensure only the best model was used. Specifically for SMAGDI, MAGDI, and SKD, 7 epochs were used for 1B and 3B models on the Strategy QA dataset, while four epochs were used for the MMLU dataset. In fine-tuning, we used 1 epoch for both datasets. We initialized the models during testing with basic prompting that informed them of their distilled roles.

Similarly, for MAGDi and SKD, the same learning rate was used and the same epochs to ensure uniformity among all models. Additionally, for MAGDi, the same parameters that were used in their original code were reused in this test. CoT prompting was used for SKD, while MAGDi had no prompting per the original code. We utilized the same MAGs across all tests, and during MAGDi testing, the decomposer and solver examples were ignored, but both pos/neg examples were created identically.

The decomposer examples were created synthetically with the prompt: "Decompose the following question into a sequence of simpler sub-questions that, when answered, " "would help solve the main question: question, based on the agent's responses" The solver examples were created synthetically with the prompt: "Answer the decompositions similar to the agent's responses"

For the fine-tuning, all LoRA configurations were set to rank: 64, alpha: 16, and dropout: 0.05.

All of our tests were run on one H200 SXM GPU for a total of 150 GPU hours.

Additionally, all results presented are from a single run.

B Datasets

For training on StrategyQA, there were 1832 training examples and 458 test examples(80-20 split). For the MMLU dataset, a subset of 1000 examples was taken from the auxiliary train. For testing, the entire 14000 test split was used. The seed of 42 was used for all splits made.

C Licenses

- StrategyQA: MIT License: License
- MMLU: MIT License: License
- MAGDi: MIT License: License
- Llama 3: MIT License: License

All information was used as per the above licenses.

D SMAGDi Responses

D.1 Initial Question (StrategyQA)

Karachi was a part of Alexander the Great's success?

Decomposition: 1. What do we mean by Alexander the Great's success? Alexander's success: Conquering from Greece to India, key wins (e.g., Gaugamela, Hydaspes), cultural spread, and empire-building via logistics and morale.

Decomposition 2: What was Alexander's involvement in the region around modern-day Karachi? He campaigned in Sindh (325 BCE), subduing tribes; his fleet used the Indus mouth (near Karachi's precursor) as a departure point for return to Persia.

Decomposition 3: Did the Karachi region contribute directly to Alexander's military victories or empire-building? Minimally direct, but yes for logistics: The harbor enabled Nearchus's fleet to map routes and return safely, supporting overall campaign closure without total loss.

Decomposition 4: Could the Karachi area have had indirect importance, like strategic or cultural value? Yes, natural harbor aided trade exploration; Greek influence in Sindh boosted later Hellenistic ties and Silk Road precursors, extending his cultural success.

Decomposition 5: So, was Karachi an important part of his success overall? Answer: True

D.2 Initial Question (MMLU)

- **Statement 1:** If H and K are subgroups of a group G, and one of H or K is a normal subgroup of G, then the set $HK = \{hk \mid h \in H, k \in K\}$ is a subgroup of G.
- **Statement 2:** All groups of order p^2, where p is a prime number, are Abelian.

These statements test key concepts in group theory: subgroup products and the structure of finite groups.

Decomposition 1: What are the two statements, and what do they claim?

- **Statement 1** claims that the product HK forms a subgroup if at least one of the subgroups H or K is normal in G.
- **Statement 2** claims that any group of order p^2 (where p is a prime) is necessarily Abelian.

Decomposition 2: Is Statement 1 true? What evidence supports this? Yes, Statement 1 is true. In group theory, if H and K are subgroups of G and at least one of them (say H) is normal in G, then HK is indeed a subgroup of G. Consider $hk, h'k' \in HK$. Then:

$$(hk)(h'k') = h(kh'k^{-1})(kk')$$

Since H is normal, $kh'k^{-1} \in H$, and hence the product is in HK. This validates closure. Without normality, closure may fail. For instance, in S_3, let:

$$A = \{e, (12)\}, \quad B = \{e, (13)\}$$

Then AB is not a subgroup, as it's not closed under composition.

Decomposition 3: What counterarguments or edge cases might challenge Statement 1? Trivial edge cases (e.g., $H = \{e\}$) still satisfy the condition. If both H and K are normal, HK is trivially a subgroup. The only potential challenge is if neither is normal, but the statement only requires one to be normal. So the statement holds.

Decomposition 4: Is Statement 2 true? What evidence supports this? Yes, Statement 2 is true. Groups of order p^2 are always Abelian. There are only two types of such groups (up to isomorphism):

$$\mathbb{Z}_{p^2}, \quad \mathbb{Z}_p \times \mathbb{Z}_p$$

Both are Abelian. By group theory results, any group of order p^2 has a nontrivial center $Z(G)$, and $|Z(G)|$ is either p or p^2. In both cases, $G/Z(G)$ is cyclic, implying G is Abelian.

Decomposition 5: What counterarguments or edge cases might challenge Statement 2? For $p = 2$, the groups of order 4 are $\mathbb{Z}_4$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$, both Abelian. For $p = 3$, groups of order 9 are also Abelian. The first non-Abelian examples occur at order p^3, such as the Heisenberg group modulo p. Thus, no counterexamples exist for order p^2.

Decomposition 6: Overall, which choice is correct based on the decomposition? Both statements are true, so the answer is True, True. Answer: 0

E SKD Responses

E.1 Initial Question (StrategyQA)

Karachi was a part of Alexander the Great's success?

1. The claim is that Karachi was a part of Alexander the Great's success.
2. Alexander the Great conquered regions around modern-day Pakistan, but his campaigns were focused mainly on cities along the Indus River.
3. Karachi is located on the Arabian Sea coast, which was not part of Alexander's main route of conquest.
4. Since Karachi was not specifically listed as one of the cities he took, it cannot have played any role in his campaigns.
5. Therefore, Karachi was not connected to Alexander the Great's success.
6. Answer: False

E.2 Initial Question (MMLU)

- **Statement 1:** If H and K are subgroups of a group G, and one of H or K is a normal subgroup of G, then the set $HK = \{hk \mid h \in H, k \in K\}$ is a subgroup of G.
- **Statement 2:** All groups of order p^2, where p is a prime number, are Abelian.

These statements test key concepts in group theory: subgroup products and the structure of finite groups.

1. Statement 1 says that if one of H or K is normal, then HK is a subgroup.
2. Consider S_3 with $H = \langle(12)\rangle$ and $K = \langle(13)\rangle$. Then HK has four elements.
3. By Lagrange's Theorem, the order of a subgroup must divide the group order (6). Since $4 \nmid 6$, HK cannot be a subgroup.
4. Therefore, this provides a counterexample, so Statement 1 is false.
5. Now consider Statement 2, which claims all groups of order p^2 are abelian.
6. However, I know that there are nonabelian p-groups such as the quaternion group of order 8.
7. Since $8 = 2^3$ is a prime power, it suggests that there can also be nonabelian groups of order p^2.
8. Therefore, not all groups of order p^2 are abelian, so Statement 2 is also false.
9. Both statements are false. Answer: 1

F Prompts

F.1 Prompts for MAS:

F.1.1 Scientist

- Generate two conflicting hypotheses before selecting an option
- Conduct a Red Team analysis attacking your own conclusion
- Calculate Bayesian probabilities for competing explanations using Bayes' theorem: $P(H \mid E) = \frac{P(E|H) \cdot P(H)}{P(E)}$
- Model system interactions using both linear and chaotic frameworks
- Compare findings against contradictory studies from adjacent fields
- Test your reasoning by asking "what could prove this wrong?"
- Consider environmental and health impacts spanning 50+ years
- Demand evidence with statistical significance before accepting claims
- Make Decision based on this

F.1.2 Lawyer

- Analyze under Common Law and Civil Law frameworks
- Simulate arguments from plaintiff/defendant perspectives simultaneously
- Identify conflicting precedents across federal circuits
- Apply game theory to predict settlement likelihoods using Nash equilibrium
- Check legality under local, national, and international law
- Identify who could sue whom if this decision is made
- Consider precedent this sets for future similar cases
- Evaluate enforceability and compliance mechanisms
- Assess constitutional and human rights implications
- Make Decision based on this

F.1.3 Historian

- Contextualize the issue within relevant historical periods and events
- Identify historical precedents and analogues for each option
- Analyze the long-term consequences of similar decisions in the past
- Examine the roles of key actors, institutions, and social forces in shaping outcomes
- Assess the reliability and biases of historical sources and narratives
- Consider the impact of cultural, economic, and technological changes over time
- Highlight lessons learned from both successes and failures in history
- Address how collective memory and historiography influence present choices
- Make Decision based on this

F.1.4 Mathematician

- Solve using both frequentist and Bayesian approaches
- Model with Monte Carlo and deterministic simulations
- Calculate error propagation through all estimation steps
- Apply robust optimization against adversarial inputs
- Quantify all variables and assign numerical values
- Calculate expected outcomes using probability theory
- Model best-case, worst-case, and most-likely scenarios
- Identify optimization targets and constraints
- Express uncertainty using confidence intervals
- Make Decision based on this

F.1.5 Ethicist

- Apply in sequence: Utilitarian, Deontological, Virtue Ethics lenses
- Calculate moral weightings using differentiable ethics equations
- Identify irreconcilable value conflicts through geometric mean analysis
- Apply multiple ethical tests: "Is this fair?", "Does this reduce suffering?", "Would I want this if roles were reversed?"
- Consider moral obligations to future generations
- Weigh individual rights against collective good
- Identify moral dilemmas and tragic trade-offs
- Question the moral legitimacy of the decision-makers
- Perform universalizability tests for proposed actions
- Make Decision based on this

F.2 Prompting for SMAGDi

F.2.1 Decomposer

Break this down into sub-questions that will help determine the answer

F.2.2 Solver

Provide a clear answer that aids in determining the answer to the main question.

F.3 COT Prompting(SKD/Fine-tuning)

You are an expert reasoning assistant. Your task is to answer True/False questions with careful analysis. Question: {question}

Instructions:

1. Let's think step by step about this question
2. Break down the key components and requirements
3. Consider what knowledge is needed to answer this
4. Apply logical reasoning to reach a conclusion
5. State your final answer as "{options}"

Analysis and Answer:

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our methodology and results prove our main claims in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the limitations of our methodology accurately in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We do not use theoretical results in our paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer:[Yes]

   Justification: We discuss all the parameters and information necessary to replicate our experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We included our data and code with open access and instructions to reproduce the main experimental results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all data splits and hyperparameters necessary to interpret results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We use statistically significant testing metrics but there are no error bars because the results were only recorded from one test.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the GPU and the runtime hours in the testing set up subsection of the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms with all aspects of the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no risks as it is based on already pretrained Llama models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: They are properly credited in the Licenses subsection of the Appendix.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide all information needed for our distillation method(asset) to be preformed.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing was used in this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: There were no participants studied in any part of this paper.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Our core methodology involves the distillation of LLM's.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.