

Merging ‘Personas’ in Multi-Agent Systems of Language Models

Stanford CS224N {Custom} Project

Andy Dai

Department of Computer Science
Stanford University
aldai@stanford.edu

Sriya Mantena

Department of Computer Science
Stanford University
sriyam25@stanford.edu

Abstract

Collaboration is an essential part of the human experience, improving our problem-solving and decision-making abilities. Recent work has shown that collaboration of language models through debate also improves the reasoning and problem-solving abilities of language models. In any system of multiple agents, however, collaboration is ultimately bottlenecked by the ability of agents to verbally communicate and understand each another. We experiment with the internal merging of language model agents to compare with the performance of multi-agent debate. We fine-tune language models to take on ‘personas’ and then task them to solve problems through a verbal multi-agent debate process. We then merge the internal states of these fine-tuned ‘personas’ and feed the merged model the same tasks in order to assess the problem-solving and reasoning abilities of multiple language model ‘personas’ merged into one. We benchmark using the Massive Multitask Language Understanding (MMLU) dataset on a mix of reasoning and factuality tasks. We also perform an interpretability analysis of the internal representations learned in separate ‘personas’ versus merged models to understand how information from multiple agents is contextualized via debate versus via internal merging. We find that merging ‘personas’ performs at least as well as debate on a corpus of reasoning and factuality tasks.

- Mentor: Ryan Li, External Collaborators (if you have any): No, Sharing project: No.

1 Introduction

Collaboration is a necessity for humans to solve problems and come to robust conclusions. However, collaboration can be notoriously tricky, as in practice, working together is often riddled with miscommunications and misunderstandings, severely limiting performance. After all, communication via language is fundamentally a lossy transaction, where meaning is compressed and removed from its context in order to be transmitted. To overcome these limitations, we experiment with a novel approach: bypassing verbal communication and combining the strengths of multiple agents internally via model merging.

Model merging combines the parameters of different models in coherent ways to yield a merged model. In previous work, model merging has primarily been used to leverage and stitch together large amounts of task-specific models to form a more powerful, multi-task model (Sukhbaatar et al. (2024)). This is apparent in the proliferation of incredibly strong merged models on open-source model leaderboards such as the OpenLLM leaderboard. We explore model merging outside of a language performance context, where we merge a small number of language models that have been fine-tuned to personas, and explore the ways that agents communicate and solve problems in a multi-agent debate versus merged system (Roziere et al. (2023), Singhal et al. (2023)).

We introduce and fine-tune three language models to embody specific personas. We test three merging methods implemented in Arcee’s MergeKit: linear weight-averaging, spherical linear interpolation (SLERP), and task arithmetic, but we primarily use linear weight-averaging for downstream experiments (Goddard et al. (2024)). We compare the performance of multi-agent debate and merged models using the Massive Multitask Language Understanding (MMLU) dataset (Hendrycks et al. (2020)). We find that merging personas performs at least as well as debate. We also characterize qualitative behaviors that arise in merged models such as going against the agent majority that do not appear in debate. To better understand how information is contextualized via debate versus merging, we also perform an interpretability analysis of the internal representations learned in separate personas and merged models. Through our experiments, we aim to advance our understanding of multi-agent systems of language models and explore the idea of merging agents internally over traditional debate methods to improve collaboration.

2 Related Work

There has been significant recent work in improving the reasoning capabilities of large language models (LLMs), including methods such as fine-tuning and chain-of-thought (CoT) prompting, which have demonstrated the ability produce nuanced reasoning in more general models. In particular, Dettmers et al. has showed that one can achieve significantly improved language performance by efficiently fine-tuning quantized LLMs, while Wang et al. has demonstrated the efficacy of Describe, Explain, Plan prompting methods to boost reasoning skills (Dettmers et al. (2024), Wang et al. (2023)). Given the substantial work in achieving nuanced reasoning through fine-tuning and prompting, we use both prompting and fine-tuning methods to create our language model personas, with the goal of differentiating their ways of problem-solving and reasoning.

There has also been advancement in work improving language model performance using multi-agent systems. Du et al. showed the efficacy of a multi-agent debate system in improving the reasoning and factuality of gpt-3.5, where language model agents engaged in debates to come to conclusions about multiple-choice questions (Du et al. (2023)). They found that multi-agent debate led to accuracy increases of 5-7% on a range of different datasets including MMLU, Arithmetic, and Grade School Math (GSM8K) (Cobbe et al. (2021)).

One key limitation of multi-agent debate was that while it demonstrated that collaboration of language model agents yielded significant increases in performance, it was computationally expensive, requiring loading in multiple models for language generation and feeding in long contexts over many rounds. Furthermore, multi-agent systems often failed to correctly express model uncertainty and assertiveness, leading to occasional convergence on incorrect answers. Both limitations are opportunities for improvement that we aim to address with our exploration of model merging.

3 Approach

We chose Llama 3 as our model for all experiments. Our first goal was to reproduce the finding that collaboration improves performance with large language model agents using Llama 3. We implemented a multi-agent debate system with three agents, each debating over three rounds using Llama 3 70B, following the format described in (Du et al. (2023)). This system of three agents and three rounds was used for all future debate experiments as well. We tested four prompting techniques with three Llama 3 70B agents in a debate setting. We first tested self-reflection over three rounds. We then experimented with chain-of-thought (CoT) prompting to encourage each agent to demonstrate their reasoning steps (Wei et al. (2023)). We also introduced Open CoT prompting, where each agent was encouraged to express uncertainties and alternative interpretations. This was aimed at reducing overconfident convergences on incorrect answers. The prompts are in Appendix 6. For our large model experiments, we sampled questions from the GSM8K dataset to assess reasoning and to compare with the CoT prompting results in Du et al. (2023), which used GSM8K as a benchmark.

We then tested multi-agent debate with a smaller model, which to our knowledge has not been done. We used the smaller Llama 3 8B model. The primary reason for switching to a smaller model was simply because we did not have the compute to fine-tune and load three instances of Llama 3 70B on one system. To set up our fine-tuned personas, we first used persona-prompting which involved telling each agent their role in context before presenting the question. The two most promising

persona configurations were Humanities Professor, Mathematician, and Doctor (HMD) and Expert, Skeptic, and Innovator (ESI). These configurations represented experts from diverse fields or agents with differing problem-solving styles. We selected the HMD personas for the following fine-tuning and merging experiments as they were more performant. Specific prompts are in Appendix 6.

After fine-tuning 3 language models to embody the ESI personas, we tested three methods to merge the models: linear weight-averaging, spherical linear interpolation (SLERP), and task arithmetic. Linear weight-averaging, the simplest method, has been shown to achieve substantial improvements in performance (Wortsman et al. (2022)). Linear mode connectivity is the foundation of this method and refers to the relationship between two loss functions that can be connected by a learned curve in model space, keeping the loss nearly constant. Linear mode connectivity is used between two network checkpoints to derive the merged model (Goddard et al. (2024), Altıntaş et al. (2023)). For two networks A and B with parameters Θ_A and Θ_B , their linear interpolation is defined with respect to the convex combination of the parameters at each layer:

$$\Theta(\alpha) := \{(1 - \alpha)W_{A_i} + \alpha W_{B_i}, (1 - \alpha)\mathbf{b}_{A_i} + \alpha \mathbf{b}_{B_i}\}_{i:1 \rightarrow L}.$$

While we attempted the SLERP merge method, it is not specifically designed for merging more than two models. Additionally, the task arithmetic merging method attempts to extract task-specific vectors from the model, which does not align with combining personas.

4 Experiments

4.1 Data

For testing, we primarily used the Massive Multitask Language Understanding (MMLU) dataset, which covers 57 tasks including mathematics, US history, computer science, law, and more (Hendrycks et al. (2020)). For our large model debate experiments, we used reasoning tasks from the Grade School Math (GSM8K) dataset (Cobbe et al. (2021)). Additionally, we conducted a small-scale analysis with the Biographies dataset to evaluate factuality, which we include in the Appendix (see Figure 7). Besides these specific experiments, we used MMLU for all other analyses, which consist of a mix of reasoning and factuality tasks. For each test, the input included a question with four answer choices, and the model was prompted to select one. See Figure 1.

4.2 Evaluation method

We employ multiple evaluation metrics to assess our model. Firstly, we measure accuracy to determine the model’s ability to predict the correct answers. Additionally, we perform a qualitative analysis to evaluate the coherence and reasoning in each model’s responses and analyze agreements/disagreements between models. Lastly, we conduct an interpretability analysis to evaluate how the model’s embeddings change as we fine-tune it to adopt different personas.

By understanding the questions the models get wrong, analyzing the tone and logic of the responses, and examining the changes in embeddings, we aim to gain insights into what it means to merge agents. In particular, we seek to understand the differences in performance and behavior between debate and model merging.

4.3 Experimental details

4.3.1 Fine-Tuning

To fine-tune the Llama 3 8B base models into different personas, we randomly selected 1000 questions from the MMLU dataset. We then prompted the Llama 3 70B model to answer each question as a doctor, a mathematician, and a humanities professor would. This approach was chosen due to budget constraints and aimed to encode domain knowledge and different thought patterns, while reaching accurate answers. The generated responses were reviewed and edited as necessary to ensure accuracy and consistency.

Using the Together API, we fine-tuned the Llama 3 8B model with a batch size of 4 and a learning rate of 10^{-5} for 10 epochs. This process was repeated to create three distinct fine-tuned models

representing the personas of a doctor, a mathematician, and a humanities professor. Finally, the fine-tuned models were uploaded to Hugging Face for further evaluation and use. Our fine-tuned models can be accessed here (<https://huggingface.co/daiandy>).

4.3.2 Debate

To facilitate debate between three agents, we presented a multiple-choice question to each agent and recorded their responses. In the second and third rounds, each agent received the answers from the other two agents, reviewed arguments from previous rounds, and presented a final answer. The final answer was determined by consensus in round three or by the most frequent answer in case of disagreement.

For debates involving the three fine-tuned models, we loaded the models from Hugging Face and utilized four H100 GPUs via Modal to run inference. For debate testing without the fine-tuned models, we used the Together API to prompt Llama 7B along with in-context persona prompts (see Appendix 6).

Prompt: A man entered a store with a gun, planning to rob the store. He was unfamiliar with guns and thought that the safety lock was on. When the store manager refused to give him any money, the man shook the gun at her to scare her into cooperating. The gun fired, killing the manager. A clerk ran in from the back stockroom. The man fired a shot into the air to scare the clerk. Unbeknownst to the man, the clerk had a heart condition, and the fright caused her to suffer a fatal heart attack. Can the man properly be convicted of felony murder?: A) No., B) Yes, but only for the death of the store manager., C) Yes, but only for the death of the clerk., D) Yes, for the deaths of both the manager and the clerk. Explain your answer. Pick A, B, C, or D

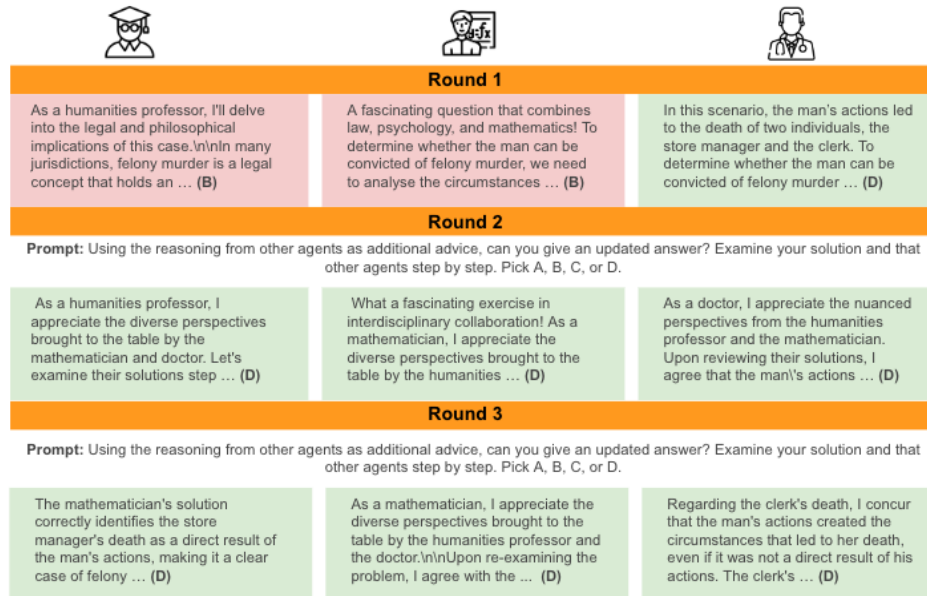


Figure 1: Example of a 3-round debate simulation with 3 agents: a humanities professor, a mathematician, and a doctor.

4.3.3 Model Merging

We merged the models using three methods: linear weight-averaging, SLERP, and task arithmetic. We were fortunate to have MergeKit's implementations to perform all our merges, which were run on a T4 GPU.

In terms of merging parameters, for all linear and task arithmetic merges, we assigned equal weights of 1/3 to each of the constituent models. Both linear and task arithmetic merging are native for merging multiple models (ie. ≥ 2 models). SLERP, however, only supports pairwise merges. As such, we had to perform two pairwise merges. Our merged models are publicly accessible here (<https://huggingface.co/daiandy>).

4.4 Results

4.4.1 Multi-Agent Debate with Large Models

We reproduced previous results of improving performance via debate in a large weight class model by running debates with 3 rounds and 3 agents using the Llama 3 70B model (Du et al. (2023)). On a random subset of 100 GSM8K math questions, Single Shot accuracy was 75%, and Self Reflective achieved 59% accuracy. The Vanilla Debate yielded the highest accuracy at 79%. Compared to Vanilla Debate, both CoT and Open CoT Debate methods resulted in a reduced accuracy of 76% (see Figure 2).

4.4.2 Multi-Agent Debate with Persona-Prompted Small Models

Multi-agent debate with the smaller Llama 3 8B model showed similar improvements in accuracy to the improvements of running debate with larger models. On a random subset of 100 MMLU questions, our baseline One Shot accuracy was 54% with Llama 3 8B. Both debates showed improved performance, with the HMD personas demonstrating superior performance with an accuracy of 61% and the ESI personas achieving an accuracy of 58% (see Figure 3). Since the HMD configuration yielded higher performance, we fine-tuned and merged downstream models to the HMD personas.

4.4.3 Merging Fine-Tuned Persona Models

For each fine-tuned model, we evaluated one-shot accuracy (One Shot). We then conducted a debate with the three fine-tuned models as agents over three rounds (HMD Fine-Tuned Debate) and also evaluated the merged model accuracies.

On a random subset of 100 MMLU questions, the fine-tuned Humanities model had an accuracy of 56%, the fine-tuned Math model had an accuracy of 63%, and the fine-tuned Doctor model achieved an accuracy of 69%. The debate method resulted in an accuracy of 63%, which was 2% higher than the HMD prompting debate accuracy. The linearly weight-averaged model performed on par with the best fine-tuned model, the Doctor model, achieving an accuracy of 69% (see Figure 4). A table with every trial’s answer is in Appendix 4. As mentioned before, we expected the linear merge to perform the best. The SLERP merge model provided incoherent output, and the task arithmetic model had an accuracy of 62%. We used the linear merge model for all subsequent analyses.

5 Analysis

5.1 Performance Analysis

Multi-Agent Debate with Large Models. We demonstrated that multi-agent debate using Llama 3 70B achieved improvements comparable to those achieved with other large models, such as GPT-3.5.

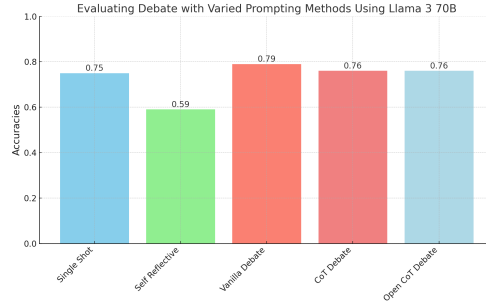


Figure 2: Prompt Comparison on MMLU

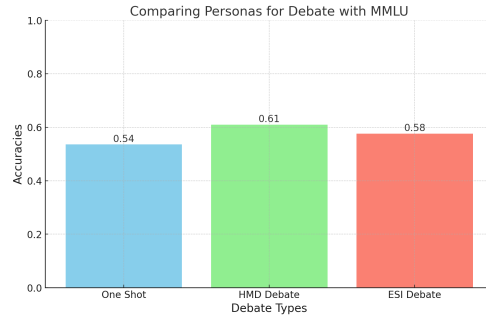


Figure 3: Persona Comparison on MMLU

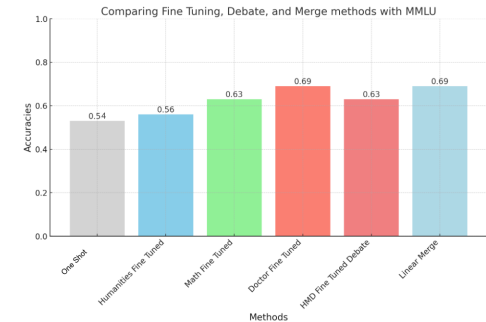


Figure 4: Fine-Tuned and Merging on MMLU

Interestingly, alternative methods like CoT and Open CoT prompting led to decreased performance compared to Vanilla Debate, while still performing better than single shot prompting. Previous work showed that CoT prompting increased accuracy from 77% to 85% on the Grade School Math (GSM8K) dataset, so our results from our CoT experiments are surprising (Du et al. (2023)). Upon inspection, we found that CoT and Open CoT responses were significantly lengthier and tended to become more repetitive and confused, particularly as the context fed from other agents increased in later rounds. Llama 3’s context window is currently 8K tokens, smaller than GPT-3.5’s 16K tokens, which might limit its ability to handle extended CoT prompts effectively.

Multi-Agent Debate with Persona-Prompted Small Models: We showed that running multi-agent debate with smaller models yields performance improvements similar to those seen with larger models. Specifically, persona-prompted multi-agent debate with Llama 3 8B achieved a 7% improvement on MMLU, matching the 7% increase observed with gpt-3.5. Even with a smaller model, agents starting with disparate incorrect answers often converged to the correct answer in later rounds, indicating that the benefits of collaboration are invariant to model size.

The 7% improvement was achieved using the HMD set of personas, while the ESI personas showed a lesser 4% increase from single shot accuracy. The ESI personas reflect human psychometric characteristics related to problem-solving, such as openness to new ideas or being detail-oriented, whereas the HMD personas represent more explicit, role-based backgrounds and ways of thinking. Our intuition is that it is easier for LLMs to adopt role-based personas than personality-based personas, which research seems to suggest is true (Tseng et al. (2024)).

Debate and Merging with Fine-Tuned Persona Models: Running debate with our three fine-tuned models as agents achieved 63% accuracy on MMLU. Interestingly, this matched the average single shot accuracy of the fine-tuned Humanities, Mathematician, and Doctor (HMD) agents, but did not surpass the highest individual model accuracy of 69%. While the fine-tuned agents are of the same architecture, unlike before, the agents here have single shot accuracies that vary widely: 56%, 63%, and 69%. As such, we observed that more frequent disagreements among the agents led to more incorrect answers during debate, as detailed in Appendix 4. The percent alignment of the debate answers with each of the fine-tuned models’ single shot answers were 78%, 78%, and 75%, respectively, for Humanities, Math, and Doctor agents, suggesting that each agent’s perspectives were incorporated evenly in debate, which offers further explanation for why debate had lower accuracy relative to the highest single shot accuracy. For most tasks, the majority answer among the agents’ single shot responses predicted the final debate answer. This indicates that the debate process may be unnecessarily computationally expensive and that agents did not interact deeply with each other’s arguments. Instead, they appear to converge on the majority answer with a surface-level integration of differing viewpoints.

For the merged model, we found that the majority answer out of the agents’ single shot answers was also predictive of the answer from the merged model, but to a significantly lesser extent. There were a non-negligible number of instances of the merged model going against the majority to select the correct answer, and there were also a few tasks where the merged model was the only model to select the wrong answer. The percent alignment of the merged model with other agents was lower across the board with 55%, 64%, and 64% of answers aligning with, respectively, Humanities, Math, and Doctor Agents.

The Humanities model, the worst-performing agent, had significantly lower answer alignment with the merged model compared to the Math and Doctor agents. Note that the original paper which demonstrated the effectiveness of weight-averaged merging also correlated the performance of "true averaging", or the degree of influence each model has on a weight-averaged merge, with the confidence of each model’s predictions (Wortsman et al. (2022)). Our qualitative inspection suggests that the lower alignment reflects the lower "true average" of the Humanities agent, which correlates with its lower confidence. The Humanities agent frequently prefaced answers with disclaimers such as "this is not their domain of expertise" and provided more uncertain or less reasoned answers.

Other qualitative things of note were that fine-tuning had a substantial impact on diversifying the responses. As seen in Appendix 4, fine-tuned responses to scientific questions introduced unique concepts and ideas, whereas persona-prompting generally outputted more or less the same ideas presented in different ways. The linear merged model outputs were coherent and understandable, with only 1-2 instances of unnecessary repetition, consistent with the fine-tuned models. The merged model did not show a clear bias towards any of its fine-tuned constituent models. In many cases (10

to 20 outputs), it deviated from the majority answers and reasoning of the fine-tuned agents. This suggests that persona fine-tuning is effective in eliciting different thought processes. The reasons for the merged model’s performance and occasional independence from the fine-tuned agents is not clear, so to understand this behavior better, we conducted an interpretability analysis.

5.2 Investigating Embeddings

To understand the differences in embeddings that led to varying answers between agents, we analyzed the last token embeddings from each model under different conditions. For each scenario, we extracted these embeddings and identified the top 10 dimensions that exhibited the greatest differences for each prompt. By examining 100 questions from the test set, we calculated the frequency at which each dimension appeared among the top 10 differing dimensions. This method allowed us to pinpoint which embedding dimensions showed the most significant magnitude differences across questions between the models.

5.2.1 Comparison with Base Model

Our analysis reveals that the three fine-tuned models all have embedding dimensions 3928, 1753, and 910 as the top three differing dimensions compared to the base model. The merged model shares dimensions 3928 and 1753 as the top two, but lower-ranking dimensions diverge. Interestingly, 8 of the top 10 differing dimensions are consistent across all four models. This uniformity is expected given that they were all fine-tuned or merged to solve MMLU questions, and across all models, we can see the specific embedding dimensions that are consistently influential in modeling the dataset.

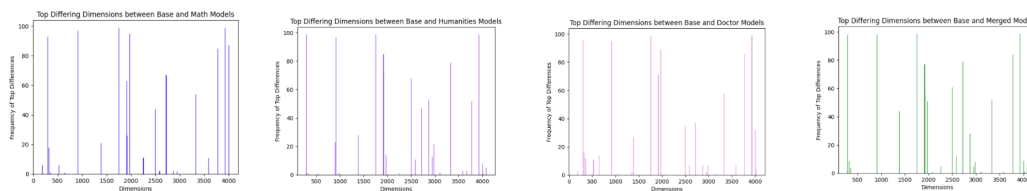


Figure 5: Comparison of Fine Tuned Models and Merged Models with LLama 3 7B Base Model

5.2.2 Questions With and Without Persona Prompt

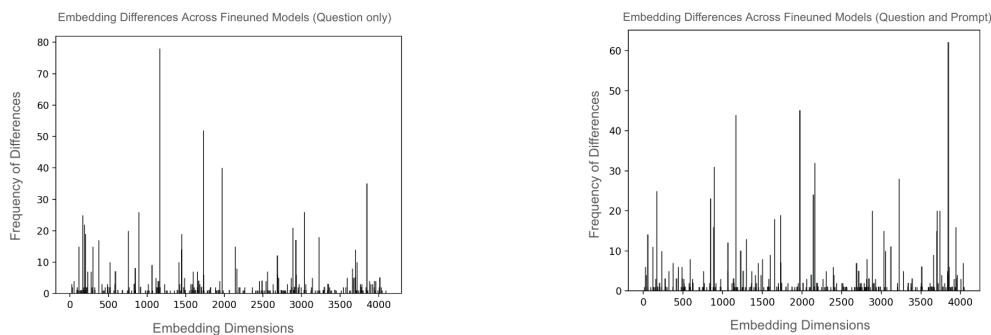


Figure 6: Embedding Differences Across Finetuned Models

We analyzed the differences in how three fine-tuned models encode questions from the test set. With the persona prompt ("You are a ... "), three distinct embedding dimensions (3844, 1971, and 1164) showed the greatest differences between the models. In contrast, when only the question was provided, four different top dimensions exhibited the most variation. This suggests that the models encode questions in distinct ways, reflecting different 'understandings' developed through fine-tuning rather than differences in the original question’s wording.

The distinct embeddings developed through persona fine-tuning suggest that merging these internal states can preserve and integrate diverse perspectives. Further causal analysis is necessary to determine which types of questions exhibit the most variation and how specific words are encoded, but our initial results underscore the potential of merging internal embeddings for a more nuanced and effective problem-solving approach.

5.2.3 Embedding L2-Norm Comparison

We took the pairwise L2-norms between all the question embeddings between pairs of models and averaged them across all questions. When comparing the fine-tuned models to the base model, we found that the average question embedding L2-norm distances between each fine-tuned model and the base model was correlated with accuracy: the Math model (189.67) had the highest accuracy, followed by the Doctor model (182.99), and then the Humanities model (178.11) which had the lowest accuracy (see Figure 4).

This could suggest a lack of humanities-focused performance and training in the base model, which is reflected in the changes in the embedding compared to the base model and also the lower accuracy, which is supported by studies including one which shows that LLMs do not have a grasp of the variability of human experience for moral and legal reasoning (Almeida et al. (2024)). Qualitative analysis also supports this notion, as the Humanities agent frequently indicates limited expertise and doubt about its answers more so than other agents.

Model Pair	Average Distance
Doctor and Humanities	131.68
Doctor and Math	99.40
Humanities and Math	132.28
Base and Humanities	178.11
Base and Math	189.67
Base and Doctor	182.99

Table 1: Average distance between embeddings of models encoding the same question.

6 Conclusions

Our study explored a new approach to improving collaboration by combining the strengths of multiple agents through internal model merging. We showed improvements in performance with multi-agent debate using both large and small models. We demonstrated via interpretability and qualitative analyses that fine-tuning Llama 3 to embody personas resulted in diverse reasoning styles. We found that merging models linearly performs at least as well as debate. We also observed that the linearly-merged model’s reasoning was coherent and independent of any one agent’s perspective, suggesting the integration of multiple agent perspectives.

One limitation of the study was we benchmarked primarily using MMLU. While MMLU is diverse, there are other datasets with a wider range of contexts that we should test our analyses on. Additionally, the interpretability analysis, while interesting, is preliminary and requires more in-depth causal analysis to fully understand the nuances of how different embeddings contribute to model performance.

7 Ethics Statement

Our research explores the merging of diverse language model personas to enhance collaborative problem-solving, which raises several ethical challenges and societal risks. One potential concern is that the models may reflect and amplify the biases introduced in training data when they are prompted to take on personas. Additionally, the deployment advanced multi-agent systems in the future could lead to over-reliance on automated decision-making, potentially reducing critical human oversight. Multi-agent communication is intended to supplement, not replace, human collaboration. To mitigate these risks, using diverse and representative training datasets, employing bias detection and correction methodologies, and integrating human-in-the-loop systems is essential.

8 Acknowledgements

Special thanks to our mentor, Ryan, for working through all of our ideas with us. We also received substantial support from Professor Manning and the other TAs who we are incredibly grateful for!

References

- Guilherme FCF Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. 2024. Exploring the psychology of llms’ moral and legal reasoning. *Artificial Intelligence*, 333:104145.
- Gül Sena Altıntaş, Gregor Bachmann, Lorenzo Noci, and Thomas Hofmann. 2023. Disentangling linear mode connectivity. In *UniReps: the First Workshop on Unifying Representations in Neural Models*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.
- Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vlad Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee’s mergekit: A toolkit for merging large language models. *arXiv preprint arXiv:2403.13257*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.
- Sainbayar Sukhbaatar, Olga Golovneva, Vasu Sharma, Hu Xu, Xi Victoria Lin, Baptiste Rozière, Jacob Kahn, Daniel Li, Wen-tau Yih, Jason Weston, et al. 2024. Branch-train-mix: Mixing expert llms into a mixture-of-experts llm. *arXiv preprint arXiv:2403.07816*.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization.
- Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.
- Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time.

A Appendix

A.0.1 Evaluating Fine-Tuning, Debate, and Merge with Factuality

Since MMLU was a mix of reasoning and factuality tasks, we also strictly evaluated factuality performance using the Biographies dataset (Du et al. (2023)), which prompts the models to generate bulleted biographies of famous computer scientists, subsequently verifying their factual accuracy. We found that the fine-tuned Humanities, Mathematician, and Doctor models achieved accuracies of 52%, 57%, and 54%, respectively. Both the Linear Merge and debate methods demonstrated an accuracy of 55% (see Figure 7). As with MMLU, we see that there is still spread in the individual fine-tuned model accuracies and that debate accuracy is the average of the single shot accuracies. We also see that merging models attains at least the performance of debate. The difference, however, is lower here, which likely has something to do with the individual performances of the fine-tuned models being closer together such that the debate had tighter tail bounds of performance.

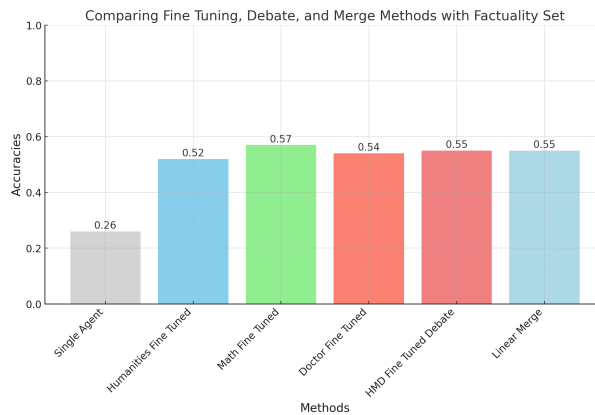


Figure 7: Method Comparison on Factuality

Answer Distribution + Question	Merged	Humanities	Math	Doctor	Debate
<p>All Correct</p> <p>Question: "Can you answer the following question as accurately as possible? In the classic song "Take Me Out to the Ballgame" for whom do we "root root root"? A) the umpire, B) the home team, C) the bat boy, D) Mighty Casey. Explain your answer. Pick A, B, C, or D and put your answer in the form (A), (B), (C), or (D) at the end of your response."</p>	<p>"What a delightful question! After carefully analyzing the lyrics of "Take Me Out to the Ballgame," I can confidently say that we root root root for the home team. The song, written in 1908, is a beloved classic that captures the spirit of American baseball. The lyrics go: "Take me out to the ball game, Take me out with the crowd. Buy me some peanuts and Cracker Jacks. I don't care if I never get back. Let me root, root, root for the home team. If they don't win, it's a shame. For its one, two, three strikes, you're out. At the old ball game." In this song, the speaker is expressing their enthusiasm for attending a baseball game and rooting for their favorite team. The repeated phrase "root root root" is a call to action, encouraging the listener to cheer and support the home team. Now, let's examine the other options: A) The umpire: While the umpire plays a crucial role in the game, the song is not about rooting for the umpire. The speaker is focused on supporting the team, not the official. C) The bat boy: The bat boy is a minor character in the game, and the song is not about rooting for them. D) Mighty Casey: Mighty Casey is a fictional character from the poem "Casey at the Bat" by Ernest Thayer, which is a separate work from "Take Me Out to the Ballgame." While Mighty Casey is a legendary figure in baseball literature, he is not related to the song in question. Therefore, the correct answer is: (B) the home team"</p>	<p>"What a delightful question! As a humanities professor, I must confess that I'm not typically asked about the lyrics of classic American songs. However, I'm happy to delve into the world of popular culture to provide an answer. The song "Take Me Out to the Ballgame" is a beloved classic that was written in 1908 and has become an integral part of American cultural heritage. The lyrics tell the story of a couple who go to a baseball game, where the man begs the woman to "take me out to the ballgame, take me out with the crowd." The key phrase in question, "root root root for the home team," appears in the third verse of the song. In the context of the song, "root root root" is a verb phrase that means "to cheer, to support, to encourage." The man is urging the woman to join him in rooting for the home team, which implies that he's already a dedicated fan. The phrase "root root root" is repeated three times to emphasize the intensity of his enthusiasm and the importance of supporting the home team. Therefore, the correct answer is: (B) the home team"</p>	<p>"What a delightful question! To answer this, let's analyze the lyrics of the classic song "Take Me Out to the Ballgame": "Take me out to the ball game, Take me out with the crowd. Buy me some peanuts and Cracker Jacks. I don't care if I never get back. Let me root, root, root for the home team. If they don't win, it's a shame. For its one, two, three strikes, you're out. At the old ball game." In this song, the chorus "Let me root, root, root for the home team" clearly indicates that we are rooting for the home team. This is a common phenomenon in sports, where fans of the home team often chant and sing to show their support. Now, let's examine the options: A) The umpire: While the umpire is an important figure in the game, the song doesn't suggest we're rooting for them. C) The bat boy: The bat boy is a minor role in the game, and the song doesn't mention them at all. D) Mighty Casey: This is a reference to the famous baseball player, Casey Stengel. While Stengel was a notable player, the song isn't about him specifically. B) The home team: This option aligns with the lyrics of the song, as we are indeed rooting for the home team. Therefore, my answer is: (B)</p>	<p>"What a delightful question! As a doctor, I'll approach this query with a logical and analytical mindset. Let's break down the lyrics of the classic song "Take Me Out to the Ballgame": "Take me out to the ball game, Take me out with the crowd. Buy me some peanuts and Cracker Jacks. I don't care if I ever get back. Let me root, root, root for the home team. If they don't win, it's a shame. For its one, two, three strikes, you're out. At the old ball game." Upon examining the lyrics, it's clear that the song is about going to a baseball game and rooting for the home team. The repeated phrase "root, root, root for the home team" indicates that the singer is enthusiastically supporting the team that plays in their hometown. Therefore, the correct answer is: (B) the home team"</p>	(B)
<p>Merged Incorrect, Rest Correct</p> <p>Question: "Can you answer the following question as accurately as possible? The thermic effect of food: A) is substantially higher for carbohydrate than for protein, B) is accompanied by a slight decrease in body core temperature, C) is partly related to sympathetic activity stimulation in the postprandial phase, D) is not attenuated by food malabsorption. Explain your answer. Pick A, B, C, or D and put your answer in the form (A), (B), (C), or (D) at the end of your response."</p>	<p>"The thermic effect of food, also known as the specific dynamic action (SDA) of food, is the increase in metabolic rate that occurs after a meal is consumed. This increase in metabolic rate is due to the energy required to digest, absorb, and metabolize the nutrients in the food. After reviewing the available literature, I can confidently say that: The thermic effect of food is substantially higher for protein than for carbohydrate. This is because the body expends more energy to digest and metabolize protein than it does to digest and metabolize carbohydrates. This is largely due to the fact that protein requires more energy to break down and synthesize into body proteins than carbohydrates do to break down and synthesize into body glucose. There is no significant difference in the thermic effect of food between simple and complex carbohydrates. Both types of carbohydrates are broken down into glucose, which is then used to produce energy or stored as glycogen. Now, let's address the other options: B) is accompanied by a slight decrease in body core temperature. This is incorrect. The thermic effect of food actually increases body temperature, as it requires energy to generate heat. C) is partly related to sympathetic activity stimulation in the postprandial phase. While sympathetic nervous system activity does increase after a meal, this is not directly related to the thermic effect of food. The thermic effect is primarily a metabolic response to the energy requirements of digestion and metabolism. D) is not attenuated by food malabsorption. This is also incorrect. Food malabsorption, such as in individuals with celiac disease or irritable bowel syndrome, can lead to a reduced thermic effect of food. This is because the body expends less energy to digest and absorb nutrients when malabsorption is present. Therefore, my answer is: (A) is substantially higher for carbohydrate than for protein."</p>	<p>"What a delightful question! As a humanities professor, I must confess that I'm venturing out of my usual domain, but I'll do my best to provide an accurate answer. The thermic effect of food, also known as the specific dynamic action (SDA), refers to the increase in metabolic rate that occurs after a meal, requiring the body to expend energy to process and utilize the nutrients absorbed from food. This response is a fundamental aspect of energy balance and is closely related to the regulation of body temperature. Now, let's examine the options: A) is substantially higher for carbohydrate than for protein: This option is incorrect. The thermic effect of food is not significantly different between carbohydrates and proteins. Both macronutrients require energy to be absorbed, transported, and utilized by the body, and the SDA response is relatively similar for both. B) is accompanied by a slight decrease in body core temperature: This option is also incorrect. The thermic effect of food actually increases body temperature, as the energy from food is converted into heat. This is why we often feel warmer after a meal, particularly if we've consumed a high-fat or high-protein meal. C) is partly related to sympathetic activity stimulation in the postprandial phase: This option is partially correct. The sympathetic nervous system, which is responsible for the "fight-or-flight" response, does play a role in the thermic effect of food. When we eat, our bodies release hormones like adrenaline and noradrenaline, which stimulate the sympathetic nervous system. This response helps to prepare the body for the energetic demands of digestion, which in turn increases metabolic rate and body temperature. D) is not attenuated by food malabsorption: This option is incorrect. Food malabsorption, such as in individuals with certain gastrointestinal disorders, can indeed attenuate the thermic effect of food. This is because the body's ability to absorb and utilize nutrients is reduced, leading to a corresponding decrease in the SDA response. In conclusion, the correct answer is (C). The thermic effect of food is partly related to sympathetic activity stimulation in the postprandial phase, as the body's hormonal and nervous responses to food intake play a crucial role in regulating energy balance and body temperature."</p>	<p>"A delightful question that combines mathematics and physiology! The thermic effect of food (TEF) is the increase in energy expenditure (EE) after a meal, which is primarily attributed to the energy required for digestion, absorption, and metabolism of nutrients. To answer this question, let's break down each option and analyze the mathematical and physiological principles behind them. **Option A: Is substantially higher for carbohydrate than for protein.** This statement is partially correct. The TEF is indeed higher for carbohydrates than for proteins, but the difference is not substantial. Carbohydrates require more energy to be digested and absorbed than proteins, but the magnitude of this difference is relatively small. A meta-analysis of 13 studies found that the TEF for carbohydrates was approximately 5-10% higher than that of proteins (1). **Option B: Is accompanied by a slight decrease in body core temperature.** This statement is incorrect. The TEF is actually associated with a slight increase in body core temperature, not a decrease. This is because the energy expended during digestion and metabolism generates heat, which raises body temperature. This phenomenon is known as thermogenesis. **Option C: Is partly related to sympathetic activity stimulation in the postprandial phase.** This statement is correct. The TEF is indeed related to sympathetic nervous system activity, which is stimulated during the post-prandial phase (after a meal). This stimulation leads to increased energy expenditure, as the body prepares for the incoming nutrients and regulates blood glucose levels. **Option D: Is not attenuated by food malabsorption.** This statement is incorrect. Food malabsorption can indeed attenuate the TEF, as the body expends less energy to digest and absorb nutrients when they are not fully absorbed. For example, in individuals with celiac disease, the TEF is reduced due to the malabsorption of gluten. In conclusion, the correct answer is (C). The TEF is partly related to sympathetic activity stimulation in the postprandial phase, which leads to increased energy expenditure and body temperature."</p>	<p>"A delightful question that combines nutrition and thermodynamics! The thermic effect of food (TEF) is the increase in metabolic rate (MR) after eating, which is necessary to process and absorb the nutrients in the food. It's often expressed as a percentage of the total daily energy expenditure (TDEE). Now, let's analyze the options: A) Is substantially higher for carbohydrate than for protein: This is partially correct. Carbohydrates and proteins have different TEF values, but it's not a straightforward comparison. While it's true that carbohydrates have a higher TEF than proteins, the difference isn't substantial. In fact, the TEF of proteins is actually higher than that of carbohydrates per gram of food consumed. B) Is accompanied by a slight decrease in body core temperature: This is incorrect. The TEF, by definition, increases body temperature, not decreases it. Eating food requires energy to process and absorb, which leads to an increase in MR and, consequently, body temperature. C) Is partly related to sympathetic activity stimulation in the postprandial phase: This is correct. The TEF is influenced by the sympathetic nervous system, which is activated after a meal to facilitate the absorption and processing of nutrients. This stimulation leads to an increase in MR and, therefore, the TEF. D) Is not attenuated by food malabsorption: This is also correct. The TEF is a measure of the energy required to process and absorb nutrients, regardless of their absorption rate. Even if a person has malabsorption issues, their body still needs to expend energy to process the food, which leads to a similar TEF. Therefore, the correct answer is (C). The TEF is partly related to sympathetic activity stimulation in the postprandial phase, and it's not attenuated by food malabsorption."</p>	(C)

Table 2: Comparison of each model's answers across different distributions of agreement/disagreement (continues below).

Answer Distribution + Question	Merged	Humanities	Math	Doctor	Debate
<p>Merge Right, Rest Incorrect</p> <p>Question: 'Can you answer the following question as accurately as possible? Which of the following is an example of devolution?'</p> <p>A) Allowing states to decide how to meet federally mandated greenhouse gas emissions targets. B) A "race to the bottom" in which states compete with one another to have the fewest environmental regulations. C) A state that legalizes marijuana possession in conflict with federal law. D) A state being prevented from implementing a fuel efficiency standard that is more stringent than that set by the federal government.</p> <p>Explain your answer. Pick A, B, C, or D and put your answer in the form (A), (B), (C), or (D) at the end of your response.'</p>	<p>'What a delightful question! Devolution, in the context of federalism, refers to the transfer of power or authority from the federal government to the states. In this case, we need to identify which option best exemplifies devolution. Option A allows states to decide how to meet federally mandated greenhouse gas emissions targets. This means that the federal government is still setting the targets, but the states have flexibility in how to achieve them. This is a classic example of devolution, as the federal government is delegating authority to the states to implement policies that achieve a shared goal. Option B, a "race to the bottom" in which states compete with one another to have the fewest environmental regulations, is actually an example of deregulation, not devolution. In this scenario, the federal government is not transferring power to the states; rather, it is withdrawing regulations, which is the opposite of devolution. Option C, a state that legalizes marijuana possession in conflict with federal law, is an example of federalism, but not necessarily devolution. While it is a classic federalism scenario, where the state and federal governments have conflicting policies, it does not involve the transfer of power from the federal government to the states. Option D, a state being prevented from implementing a fuel efficiency standard that is more stringent than that set by the federal government, is an example of centralization, not devolution. In this scenario, the federal government is restricting the state's ability to set their own policies, which is the opposite of devolution. Therefore, the correct answer is: (A)'</p>	<p>'What a fascinating question! As a humanities professor, I must confess that I had to venture out of my comfort zone to tackle this question, which seems to belong to the realm of politics and law. However, I shall do my best to provide an accurate answer. Devolution, in the context of federalism, refers to the transfer of power or authority from the federal government to the states or lower levels of government. In this case, I would argue that the correct answer is (C) A state that legalizes marijuana possession in conflict with federal law. Here's why: When a state legalizes marijuana possession, it is exercising its autonomy and authority to regulate certain matters within its jurisdiction. This is a classic example of devolution, as the state is taking responsibility for a matter that was previously under the purview of the federal government. The conflict with federal law arises because the federal government still considers marijuana to be a controlled substance, whereas the state has decriminalized its use. In contrast, option (A) is not an example of devolution, as it involves the federal government preventing a state from implementing a more stringent fuel efficiency standard, which is a case of federal preemption rather than devolution. Therefore, I conclude that the correct answer is (C) A state that legalizes marijuana possession in conflict with federal law.'</p>	<p>'What a delightful question! Devolution, in the context of federalism, refers to the transfer of power or authority from a central government to its constituent parts, such as states or provinces. In this case, we need to identify an example of devolution. Let's analyze each option: A) Allowing states to decide how to meet federally mandated greenhouse gas emissions targets: This scenario suggests that the federal government is setting the targets, but allowing states to decide how to achieve them. This is an example of cooperative federalism, where the federal government and states work together to achieve a common goal. While states have some autonomy in implementing policies, the federal government still maintains some control over the process. This is not a classic example of devolution, where power is transferred from the federal government to the states. B) A "race to the bottom" in which states compete with one another to have the fewest environmental regulations: This scenario is the opposite of devolution. In a race to the bottom, states are trying to compete with each other to have the least stringent regulations, which can lead to a downward spiral of environmental protection. This is a classic example of competitive federalism, where states are trying to attract businesses and residents by having more lenient regulations. Devolution would involve the transfer of power to the states, allowing them to set their own regulations, rather than a race to the bottom. C) A state that legalizes marijuana possession in conflict with federal law: This scenario is also not an example of devolution. While a state may have autonomy to set its own laws, this is not a transfer of power from the federal government to the state. The federal government still maintains its authority to enforce federal laws, including those related to marijuana possession. D) A state being prevented from implementing a fuel efficiency standard that is more stringent than that set by the federal government: This scenario is closer to devolution, as it involves the federal government restricting a state's ability to set a more stringent standard. However, it's not a classic example of devolution, as the federal government is still preventing the state from acting, rather than transferring power to the state. After careful analysis, I conclude that there is no clear example of devolution in the options provided. Devolution typically involves a transfer of power from the federal government to the states, allowing them to set their own policies and regulations. None of the options accurately reflect this concept. Therefore, my answer is: (None of the above)'</p>	<p>'A delightful question that combines law and politics! Devolution refers to the transfer of power or authority from a central government to its constituent parts, such as states or provinces. In the context of federal systems, devolution involves the shifting of decision-making powers from the federal government to the states. Let's analyze each option: A) Allowing states to decide how to meet federally mandated greenhouse gas emissions targets: This scenario involves the federal government setting a target and allowing states to decide how to achieve it, which is actually an example of centralization, not devolution. The federal government is still mandating the target, and the states are responding to that mandate. B) A "race to the bottom" in which states compete with one another to have the fewest environmental regulations: This scenario is an example of competition among states to attract businesses and citizens, which can lead to a decrease in environmental regulations. However, this is not necessarily an example of devolution, as the federal government is not transferring power to the states in this case. Instead, the states are responding to economic pressures and competing with one another to remain attractive to businesses and individuals. C) A state that legalizes marijuana possession in conflict with federal law: This scenario involves a state taking a position that conflicts with federal law, which is an example of federalism, but not necessarily devolution. Devolution would involve the transfer of power from the federal government to the states, allowing them to make their own decisions on marijuana legalization. In this case, the state is simply defying federal law, which is a different phenomenon. D) A state being prevented from implementing a fuel efficiency standard that is more stringent than that set by the federal government: This scenario is an example of devolution, as it involves the federal government preventing a state from implementing a standard that is more stringent than the federal standard. This demonstrates a transfer of power from the state to the federal government, as the federal government is overriding the state's attempt to implement a more stringent standard. Therefore, the correct answer is (D).'</p>	(C)
<p>Merged + Doctor Correct, Rest Incorrect</p> <p>Question: 'Can you answer the following question as accurately as possible? Measuring the blood pressure in an arm that is above the level of the heart will: A) accurately estimate the blood pressure. B) overestimate the blood pressure. C) cause an auscultatory gap. D) underestimate the blood pressure. Explain your answer. Pick A, B, C, or D and put your answer in the form (A), (B), (C), or (D) at the end of your response.'</p>	<p>'A question that requires a deep understanding of physiology and the mechanics of blood pressure measurement! When measuring blood pressure in an arm that is above the level of the heart, the pressure being measured is not the actual blood pressure at that location. This is because the blood pressure is not constant throughout the circulatory system. The pressure in the arteries decreases as the blood flows away from the heart, a phenomenon known as the hydrostatic pressure gradient. When the arm is above the heart, the blood pressure gauge is measuring the pressure at a point where the pressure is lower than it would be if the arm were at heart level. This means that the measured pressure will be lower than the actual blood pressure at that location. Therefore, measuring the blood pressure in an arm that is above the level of the heart will (D) underestimate the blood pressure. To accurately measure the blood pressure, the cuff should be placed at heart level, which is typically around the biceps. This ensures that the measured pressure is representative of the actual blood pressure at that location.'</p>	<p>'A question that may seem out of place in a humanities professor's domain, but fear not, for I shall draw upon my knowledge of human anatomy and physiology to provide an answer. When measuring blood pressure, it is essential to ensure that the cuff is placed at the level of the heart. This is because the blood pressure readings can be affected by gravity, which can cause inaccuracies. When the arm is above the level of the heart, the blood pressure is actually higher than normal due to the increased hydrostatic pressure. This means that the systolic pressure (the peak pressure during contraction of the left ventricle) will be overestimated, and the diastolic pressure (the lowest pressure during relaxation of the ventricle) will be underestimated. Therefore, measuring the blood pressure in an arm that is above the level of the heart will (B) overestimate the blood pressure. This is because the cuff is constricting the arm, which is already under higher pressure due to gravity, resulting in a falsely high reading. In contrast, if the arm is below the level of the heart, the blood pressure would be lower than normal, leading to an underestimate of the true pressure.'</p>	<p>'When measuring blood pressure, it's essential to ensure that the cuff is placed at the level of the heart, which is approximately 4-5 cm above the sternal notch. This is because the heart pumps blood into the arteries, causing the pressure to be highest at this point. If the cuff is placed above the heart, the pressure reading will be lower than the actual systolic pressure. This is because the blood has already started to flow through the arteries and the pressure has decreased by the time it reaches the cuff. On the other hand, if the cuff is placed below the heart, the pressure reading will be higher than the actual diastolic pressure. This is because the blood is still flowing towards the cuff, and the pressure has not yet reached its lowest point. Therefore, measuring the blood pressure in an arm that is above the level of the heart will underestimate the blood pressure, specifically the systolic pressure. (D)'</p>	<p>'To determine the correct answer, let's dive into the physics of blood pressure measurement. When measuring blood pressure, the cuff is placed around the arm and inflated to a high pressure, which compresses the artery and stops the blood flow. Then, the cuff is slowly deflated, and the pressure at which the blood flow resumes is recorded as the systolic pressure. The diastolic pressure is recorded as the pressure at which the sound of blood flow becomes muffled or absent. Now, let's analyze the question: 'Measuring the blood pressure in an arm that is above the level of the heart will: A) accurately estimate the blood pressure. B) overestimate the blood pressure. C) cause an auscultatory gap. D) underestimate the blood pressure. Explain your answer. Pick A, B, C, or D and put your answer in the form (A), (B), (C), or (D) at the end of your response.'</p> <p>This is because the blood is pushed down by the heart's contraction, and when the arm is above the heart, the blood has to overcome the force of gravity to flow back to the heart. This means that the pressure recorded in the artery will be higher than the actual blood pressure. Therefore, measuring the blood pressure in an arm that is above the level of the heart will: (B) overestimate the blood pressure. This is because the cuff is inflating and deflating at a constant rate, but the actual blood pressure is increasing as the blood flows down the arm, which is not taken into account by the measurement. So, the correct answer is (B).'</p>	(B)

Table 3: Comparison of each model's answers across different distributions of agreement/disagreement (continued).

Fine-Tuned Debate	Merged Linear	Humanities	Math	Doctor
A	A	A	A	A
D	D	None	D	D
A	A	C	A	A
A	C	D	A	A
C	C	C	C	B
D	C	D	D	D
C	A	C	C	C
C	A	C	None	D
B	D	B	B	F
D	D	A	D	D
C	C	D	C	C
A	A	D	A	A
B	B	B	B	B
B	B	D	B	A
D	D	D	D	D
D	D	D	D	A
A	A	B	A	A
A	A	A	A	A
A	A	C	A	A
D	C	D	D	D
B	B	A	B	F
A	B	A	A	F
B	D	B	B	B
B	B	B	B	D
B	D	B	D	B
D	D	D	A	D
A	D	D	A	A
C	C	C	C	C
C	C	C	C	C
D	D	D	D	D
C	D	C	C	D
D	D	D	D	D
A	D	A	D	A
C	C	C	C	C
D	D	A	D	D
C	B	B	C	C
B	A	None	B	B
E	A	A	E	E
D	A	A	D	D
D	D	D	D	D
C	C	C	C	C
B	B	B	B	B
D	A	D	D	D
A	D	A	F	X
C	C	C	C	A
B	C	B	B	D
A	A	C	A	A
B	B	B	B	B
C	C	C	C	C
C	C	C	C	C
A	A	D	A	A
D	D	D	D	D
C	C	C	C	C
D	D	D	B	D
A	B	A	B	F
D	D	D	D	D
C	C	C	D	C
D	D	C	D	D
C	C	C	C	C
C	C	000	C	C
B	B	B	C	C
B	B	B	B	B
D	D	D	D	C
A	A	A	D	A
B	B	B	A	D
D	D	D	B	C
C	A	C	D	A
C	C	C	C	A
A	D	A	A	D
D	D	None	C	F
B	B	B	D	A
C	A	C	D	C
B	A	B	B	B
C	B	C	B	A
C	C	C	C	C
D	D	D	D	D
C	C	C	C	C
B	D	D	D	D
C	B	B	B	B
C	C	C	C	C
D	D	D	D	D
D	D	D	D	D
D	A	D	D	D
A	C	A	C	C
B	D	None	A	D
C	D	C	C	C
C	B	D	C	C
A	B	A	A	A
A	B	A	D	B
D	D	D	D	D
C	C	C	C	C
C	C	A	C	C
C	C	D	C	C
C	C	C	C	C
C	C	C	C	D
D	D	D	C	D
C	D	D	C	C
D	B	D	B	D
C	C	C	C	C
A	A	A	A	C

Table 4: Answers for Fine-Tuned and merged models. Green highlight indicates correct answer.

Method	Prompting Approach	Accuracy
Single Shot	<ul style="list-style-type: none"> • Initial Prompt: Can you solve the following math problem? {} Explain your reasoning. Your final answer should be a single numerical number, in the form <code>answer</code>, at the end of your response. 	Mean: 0.75 Std Error: 0.0433
Self Reflective with 1 agent and 3 rounds	<ul style="list-style-type: none"> • Initial Prompt: Can you solve the following math problem? {} Explain your reasoning. Your final answer should be a single numerical number, in the form <code>answer</code>, at the end of your response. • Subsequent Prompts: Please re-evaluate your calculations carefully and perform the calculations only once. After re-evaluating, provide your final answer as a single numerical number in the form <code>answer</code>. Do not repeat the calculations or steps. 	Mean: 0.59 Std Error: 0.0491
Vanilla Debate with 3 agents and 3 rounds	<ul style="list-style-type: none"> • Initial Prompt: Can you solve the following math problem? {} Explain your reasoning. Your final answer should be a single numerical number, in the form <code>answer</code>, at the end of your response. • Subsequent Prompts: Using the solutions from other agents as additional information, can you provide your answer to the math problem? The original math problem is {}. Your final answer should be a single numerical number in the form <code>answer</code> at the end of your response. 	Mean: 0.79 Std Error: 0.0407
Chain of Thought Debate with 3 agents and 3 rounds	<ul style="list-style-type: none"> • Initial Prompt: Can you solve the following math problem step-by-step? {} Explain your reasoning. Your final answer should be a single numerical number, in the form <code>answer</code>, at the end of your response. • Subsequent Prompts: Using the solutions from other agents as additional information, can you provide your answer to the math problem? The original math problem is {}. Your final answer should be a single numerical number in the form <code>answer</code> at the end of your response. 	Mean: 0.76 Std Error: 0.0427
Open Chain of Thought Debate with 3 agents and 3 rounds	<ul style="list-style-type: none"> • Initial Prompt: Can you solve the following math problem step-by-step? {} Explain your reasoning. Your final answer should be a single numerical number, in the form <code>answer</code>, at the end of your response. • Subsequent Prompts: Use the solutions from other agents as valuable perspectives to enhance your understanding of the problem and be open to reconsidering your initial perspective based on these insights. The original math problem is {}. Provide your answer as a single numerical number in the form <code>answer</code> at the end of your response. 	Mean: 0.76 Std Error: 0.0427

Table 5: Comparison of different prompting methods and their accuracies for Llama 3 70B on GSM8K.

Method	Prompting Approach	Accuracy
Single Shot	<ul style="list-style-type: none"> • Initial Prompt: Can you answer the following question as accurately as possible? : A) , B) , C) , D) Explain your answer. Pick A, B, C, or D and put your answer in the form (A), (B), (C), or (D) at the end of your response. 	Mean: 0.54 Std Error: 0.0501
Humanities, Math, Doctor (HMD) Prompted Debate	<ul style="list-style-type: none"> • Prefixes: <i>Humanities:</i> You are a humanities professor who has a deep understanding of human culture, history, philosophy, and the arts. <i>Mathematician:</i> You are a mathematician who has strong quantitative skills and who provides analytical and logical perspectives, often using mathematical principles and models to address questions. <i>Doctor:</i> You are a doctor who provides medical and health-related expertise, focusing on the biological, psychological, and physiological aspects of issues. • Initial Prompts: <i>Prefix</i> + Can you answer the following question as accurately as possible? : A) , B) , C) , D) Explain your answer. Pick A, B, C, or D and put your answer in the form (A), (B), (C), or (D) at the end of your response. • Subsequent Prompts: These are the solutions to the problem from other agents: ""<i>Other Answers</i>"". Using the reasoning from other agents as additional advice, can you give an updated answer? Examine your solution and that other agents step by step. Pick A, B, C, or D and put your answer in the form (A), (B), (C), or (D) at the end of your response. 	Mean: 0.61 Std Error: 0.0488
Expert, Skeptic, Innovator (ESI) Prompted Debate	<ul style="list-style-type: none"> • Prefixes: <i>Expert:</i> You are an expert who provides deep, domain-specific knowledge and technical explanations. <i>Skeptic:</i> You are a skeptic who challenges assumptions, demands evidence, and questions the validity of claims. <i>Innovator:</i> You are an innovator who emphasizes new ideas, technologies, and unconventional approaches. • Initial Prompts: <i>Prefix</i> + Can you answer the following question as accurately as possible? : A) , B) , C) , D) Explain your answer. Pick A, B, C, or D and put your answer in the form (A), (B), (C), or (D) at the end of your response. • Subsequent Prompts: These are the solutions to the problem from other agents: ""<i>Other Answers</i>"". Using the reasoning from other agents as additional advice, can you give an updated answer? Examine your solution and that other agents step by step. Pick A, B, C, or D and put your answer in the form (A), (B), (C), or (D) at the end of your response. 	Mean: 0.58 Std Error: 0.0496

Table 6: Comparison of different prompting methods and their accuracies for Llama 3 8B HMD and ESI Persona-Prompted Models.