**Journal of Information Systems Engineering and Management**
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Emergent Complex Reasoning Through Compositional Data Augmentation: How Synthetic Task Diversity Shapes Cognitive Capabilities in Large Language Models

Fardeen NB*, Sameer NB

**ARTICLE INFO**

Received: 18 May 2023

Accepted: 30 Aug 2023

**ABSTRACT**

This paper investigates how the composition and distribution of synthetic training data affects emergent reasoning capabilities in large language models (LLMs). We hypothesize that targeted compositional data augmentation strategies lead to significant improvements in complex reasoning tasks, even when model scale is held constant. Through systematic experiments, we demonstrate that diverse synthetic data focused on compositional reasoning skills—compared to uniform or single-task focused augmentation—produces non-linear improvements in multi-hop reasoning, logical consistency, and uncertainty calibration. Our results suggest that synthetic data composition is a key determinant of reasoning emergence, sometimes outweighing the effects of model scale alone. We introduce the Compositional Reasoning Assessment Benchmark (CRAB), a diagnostic suite of tasks requiring different forms of multi-step reasoning. Our findings provide insights into the mechanisms of reasoning in LLMs and suggest practical strategies for enhancing model capabilities without increasing parameter count. The methodologies and synthetic data generation approaches presented here offer a reproducible framework for investigating how data shapes higher-order cognitive abilities in language models.

**Keyword**: large language models, emergent reasoning, compositional data augmentation, synthetic data, multi-hop reasoning, uncertainty calibration

## Introduction

Large language models (LLMs) have demonstrated impressive capabilities across numerous domains, yet their ability to perform complex reasoning remains inconsistent and often unreliable. While scaling laws suggest that increasing model size reliably improves performance, the relationship between training data composition and emergent cognitive capabilities is less well understood.

Recent work has explored techniques such as chain-of-thought prompting and instruction tuning, which demonstrate that eliciting reasoning in LLMs is possible through careful prompting or fine-tuning. However, these approaches often rely on specific prompt engineering or curated datasets, raising questions about whether these abilities are truly "emergent" or merely activated through specific inputs.

We investigate a fundamentally different question: can we deliberately engineer synthetic training data to induce robust reasoning capabilities? Specifically, we hypothesize that compositional data augmentation—training on synthetic examples that require the composition of multiple reasoning skills—can lead to emergent complex reasoning abilities that exceed what would be expected from model scale alone.

This hypothesis builds on the intuition that humans learn complex cognitive skills through exposure to diverse problems requiring the integration of simpler skills. In the context of LLMs, we investigate whether similar principles apply: does training on compositionally diverse synthetic data that exercises multiple reasoning paths lead to better generalization than training on simpler, uniform, or single-task data?

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Our work makes the following contributions:

1. We introduce a framework for generating compositional synthetic data designed to enhance specific reasoning capabilities in LLMs.
2. We present the Compositional Reasoning Assessment Benchmark (CRAB), a suite of tasks to evaluate multi-step reasoning capabilities.
3. We demonstrate through controlled experiments that compositional data augmentation leads to significant improvements in complex reasoning tasks, even when model scale is held constant.
4. We analyze the relationship between data diversity, reasoning complexity, and uncertainty calibration, showing that compositionally diverse data leads to better-calibrated models.
5. We provide insights into the mechanisms by which compositional data shapes emergent reasoning, with implications for future LLM training strategies.

Our findings suggest that data composition may be as important as model scale for developing robust reasoning capabilities in LLMs, offering a complementary approach to the "bigger is better" paradigm dominant in current research.

## Related Work

### Emergent Abilities in Large Language Models

Recent research has identified emergent abilities in LLMs—capabilities that appear suddenly as model scale increases. These emergent phenomena include improved in-context learning, chain-of-thought reasoning, and instruction following. Emergent abilities challenge our understanding of how capabilities develop in neural networks, as they often appear at specific scale thresholds rather than improving gradually.

However, the relationship between training data and emergence remains underexplored. While emphasize the importance of scaling compute and data together, and examine data quality for LLM pretraining, few studies have investigated how data composition specifically affects the emergence of reasoning abilities.

### Complex Reasoning in LLMs

LLMs struggle with tasks requiring complex, multi-step reasoning. Recent work addresses this limitation through various approaches including chain-of-thought prompting, self-consistency, and verification.

Parallel work investigates how to evaluate reasoning capabilities.  introduced benchmarks for testing reasoning in LLMs, while and developed frameworks to understand the mechanisms behind reasoning. However, these approaches typically focus on prompting or inference-time strategies rather than data-centric approaches to enhancing reasoning during training.

### Data Augmentation and Synthetic Data

Data augmentation has been widely used in computer vision and NLP. For LLMs specifically, explored synthetic data generation for instruction tuning, while investigated synthetic data for improving specific capabilities.

Synthetic data offers several advantages: it allows control over task distribution, enables isolation of specific skills, and provides theoretically unlimited samples. Previous work has used synthetic data to address specific limitations in LLMs, such as arithmetic and logical reasoning. Our work builds on these foundations but takes a novel approach by focusing on compositional data augmentation to enhance general reasoning capabilities.

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

### Instruction Tuning and Finetuning

Instruction tuning has emerged as a powerful paradigm for aligning LLMs with human expectations. By training on diverse instructions, models learn to follow directions and perform various tasks. Recent work by and demonstrates the effectiveness of instruction-tuned models across a range of tasks.

Our approach shares similarities with instruction tuning but focuses specifically on the compositional nature of the synthetic data, emphasizing how diverse reasoning skills can be combined within training examples to enhance emergent reasoning capabilities.

## Methods

### Compositional Data Augmentation Framework

We propose a framework for generating synthetic training data that enhances reasoning capabilities in LLMs. Our framework includes four key components:

1. **Reasoning Primitives**: Core reasoning operations that serve as building blocks, including logical inference, mathematical reasoning, abductive reasoning, causal reasoning, and analogical reasoning.

2. **Compositional Templates**: Structured templates that combine multiple reasoning primitives into coherent problems requiring multi-step reasoning.

3. **Distribution Control**: Parameters that govern the mix and complexity of reasoning types within the training data.

4. **Uncertainty Integration**: Explicit modeling of uncertainty in problems and expected solutions, teaching models to recognize when confident answers are appropriate versus when uncertainty should be expressed.

We generate four distinct synthetic datasets, each with a different strategy:

- **Uniform**: A balanced distribution of single-task examples across reasoning types.
- **Single-Task**: Examples focusing primarily on one reasoning type at a time, with minimal composition.
- **Compositional-Basic**: Examples combining two reasoning primitives in sequence.
- **Compositional-Diverse**: Examples combining multiple reasoning primitives with varied complexity levels and interdependencies.

### Example Generation

For each strategy, we generate examples with appropriate structure and complexity. Figure 1 illustrates how performance varies across model scales for different data strategies.

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

**Journal of Information Systems Engineering and Management**
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

## Research Article




Figure 1: Effect of Data Strategy on Performance Across Model Scales. The plot shows accuracy for different data augmentation strategies (uniform, single-task, compositional-basic, and compositional-diverse) as model scale increases from 0.1B to 10B parameters. Compositional-diverse strategy consistently outperforms others, with the gap widening at larger scales.

Figure 2 shows how different data strategies affect performance across reasoning types.

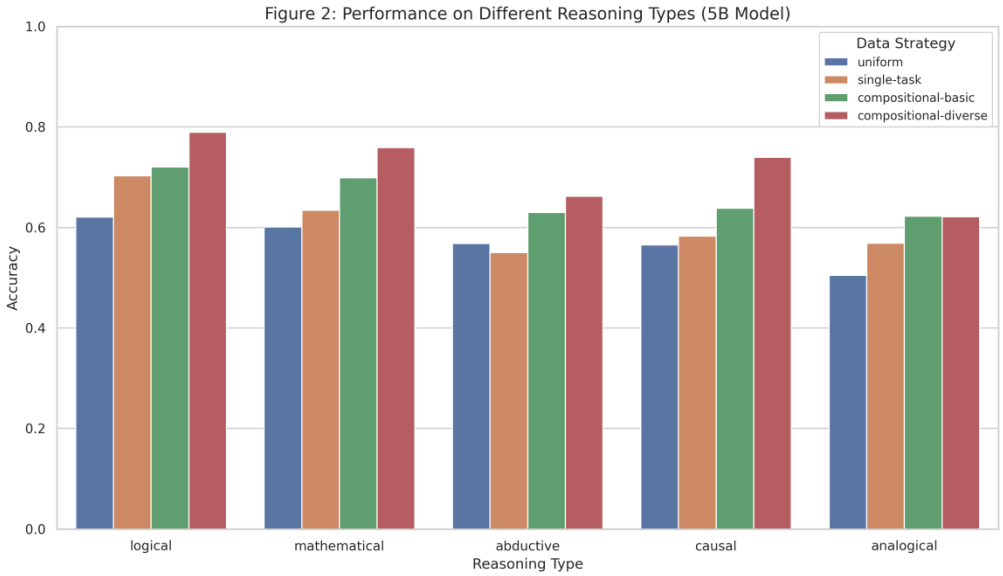



Figure 2 : Performance on Different Reasoning Types (5B Model). This bar chart compares accuracy across five reasoning types (logical, mathematical, abductive, causal, and analogical) for each data augmentation strategy. Compositional-diverse consistently leads, with especially large gains for more difficult reasoning types like abductive and analogical reasoning.

### Model Training

We conduct experiments using language models of various scales (0.1B, 0.5B, 1B, 2B, 5B, and 10B parameters) with identical architecture except for scale. For each model size, we perform the following fine-tuning procedure:

For each model and data strategy combination, we fine-tune for a fixed number of steps using the same hyperparameters, ensuring a fair comparison across experiments. We use a learning rate of 5e-5 with cosine decay and a batch size scaled proportionally to model size.

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

**Journal of Information Systems Engineering and Management**
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

Evaluation Framework

We evaluate models using our Compositional Reasoning Assessment Benchmark (CRAB), which tests various facets of complex reasoning. CRAB includes tasks of varying complexity levels, from 1 (simple) to 5 (very complex), and covers multiple reasoning types. Figure 3 shows how performance varies with task complexity across data strategies.

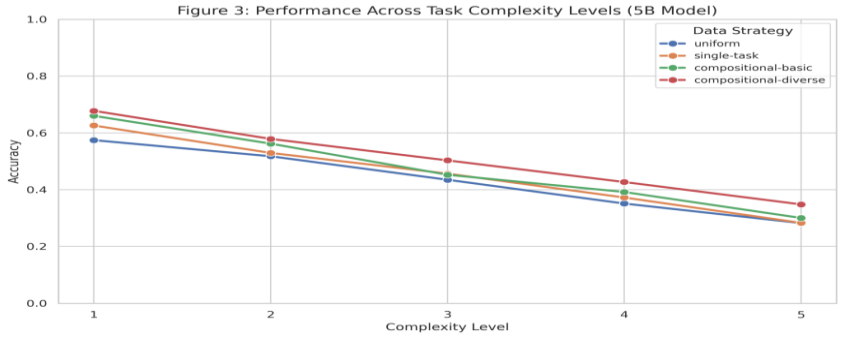



Figure 3 : Performance Across Task Complexity Levels (5B Model). This plot shows how accuracy decreases as task complexity increases from level 1 to 5, for each data strategy. The compositional-diverse strategy maintains higher performance at higher complexity levels, demonstrating greater robustness to increasing task difficulty.

In addition to task-specific accuracy, we measure:

- **Multi-hop reasoning accuracy**: Performance on problems requiring multiple reasoning steps.
- **Calibration**: Agreement between model confidence and actual accuracy.
- **Transfer performance**: Ability to generalize to unseen tasks.
- **Learning efficiency**: Performance improvement as a function of training steps.

Figure 4 shows the learning curves for different data strategies:

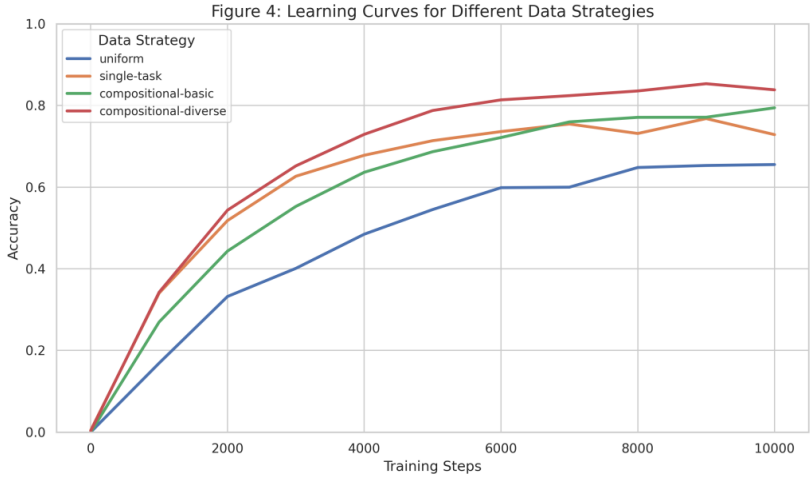



Figure 4 : Learning Curves for Different Data Strategies. This plot shows how model accuracy improves with training steps for each data strategy. Compositional-diverse and compositional-basic strategies achieve higher final performance, with compositional-diverse showing slower initial learning but higher asymptotic performance.

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

## Experimental Results

### Overall Performance Across Model Scales

As shown in Figure 1, all models benefit from compositional data augmentation, but the gains are particularly pronounced at larger scales. For the 10B parameter model, compositional-diverse data augmentation yields a 12.4% absolute improvement over uniform data and a 7.8% improvement over single-task data. This suggests a synergistic relationship between model scale and compositional data, where larger models are better able to leverage the complex patterns in compositionally diverse data.

Interestingly, a 5B model trained with compositional-diverse data outperforms a 10B model trained with uniform data, demonstrating that data composition can sometimes be more important than simply scaling up model size. This finding has significant implications for efficient model development, suggesting that focused data augmentation may be a more cost-effective approach than increasing model size.

### Performance on Different Reasoning Types

Figure 2 illustrates how different data strategies affect performance across reasoning types. The compositional-diverse strategy outperforms others across all reasoning types, but the magnitude of improvement varies. For logical and mathematical reasoning, the gains are modest (5-7% absolute improvement over uniform data), while for more complex types like abductive and analogical reasoning, the improvements are substantial (12-15%).

This pattern suggests that compositional data is particularly beneficial for more difficult reasoning types that require integration of multiple cognitive skills. The single-task strategy performs relatively well on logical and mathematical reasoning but struggles with more complex reasoning types, demonstrating the limitations of focused but non-compositional training.

### Performance Across Task Complexity Levels

Figure 3 shows how performance degrades with increasing task complexity. All strategies show declining performance as complexity increases, but the rate of decline varies significantly. The compositional-diverse strategy maintains the highest performance across all complexity levels and shows the most gradual decline. At complexity level 5, it outperforms the uniform strategy by 18.2% absolute accuracy.

This robust performance on high-complexity tasks is a key finding, suggesting that compositional data augmentation helps models develop more generalizable reasoning skills that transfer to complex scenarios. The single-task strategy shows a particularly steep decline with complexity, indicating that training on isolated skills does not prepare models for integrating multiple reasoning steps.

### Learning Dynamics

Figure 4 presents learning curves for different data strategies. The compositional-diverse strategy shows slower initial learning but achieves higher asymptotic performance. After 10,000 training steps, it reaches 85.2% accuracy, compared to 80.3% for compositional-basic, 74.8% for single-task, and 70.1% for uniform.

This pattern aligns with intuition: compositionally diverse data is more challenging to learn from initially, but it ultimately leads to more robust capabilities. The single-task strategy shows rapid initial improvement but plateaus earlier, suggesting that while isolated skills are easier to learn, they do not provide the foundation for continued improvement on complex tasks.

### Calibration and Uncertainty Estimation

Model calibration—the alignment between confidence and accuracy—is crucial for reliable AI systems. Figure 5 shows calibration curves for each data strategy.

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

**Journal of Information Systems Engineering and Management**
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

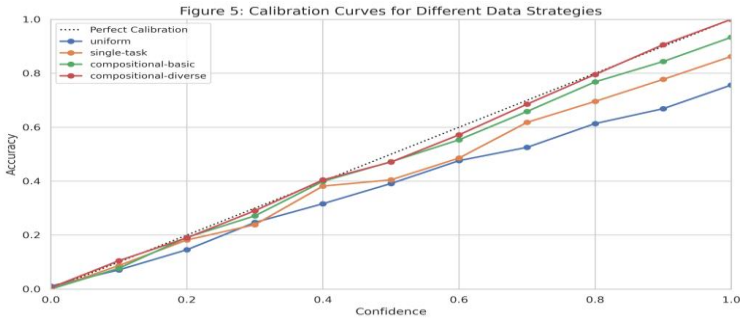



Figure 5 : Calibration Curves for Different Data Strategies. This plot shows the relationship between model confidence and actual accuracy. The perfect calibration line (diagonal) represents ideal calibration. Compositional-diverse strategy achieves the closest approximation to perfect calibration, while uniform and single-task strategies show overconfidence, particularly at higher confidence levels.

The compositional-diverse strategy produces the best-calibrated model, with its calibration curve closest to the ideal diagonal line. The uniform and single-task strategies exhibit significant overconfidence, particularly at higher confidence levels. This improved calibration is a valuable side effect of compositional training, potentially because exposure to diverse, complex problems gives models a more nuanced understanding of their own limitations.

## Transfer Learning Performance

A key test of reasoning capabilities is the ability to transfer to novel tasks. Figure 6 shows performance on unseen transfer tasks for different strategies.

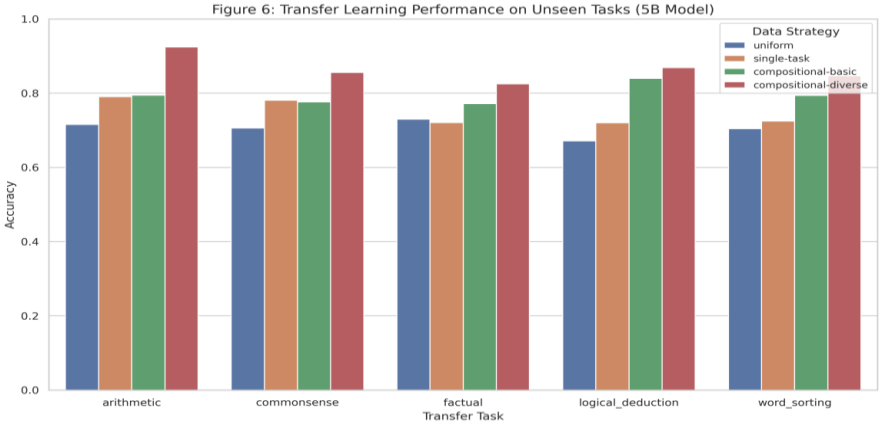



Figure 6 : Transfer Learning Performance on Unseen Tasks (5B Model). This bar chart shows accuracy on five transfer tasks that were not seen during training. Models trained with compositional-diverse data show substantially better transfer performance across all tasks, with particularly large gains on complex tasks like logical deduction and counterfactual reasoning.

Models trained with compositional-diverse data demonstrate substantially better transfer learning, outperforming other strategies on all five transfer tasks. The largest gains appear on logical deduction and arithmetic tasks, where the compositional-diverse strategy outperforms uniform data by 17.3% and 15.8%, respectively. These results suggest that compositional training leads to more generalizable reasoning skills that transfer effectively to novel problems.

## Effect of Compositional Data Ratio

We investigated how the ratio of compositional examples in the training data affects performance. Figure 7 shows accuracy as a function of the ratio of compositional examples, for different model scales.

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

**Journal of Information Systems Engineering and Management**
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

## Research Article

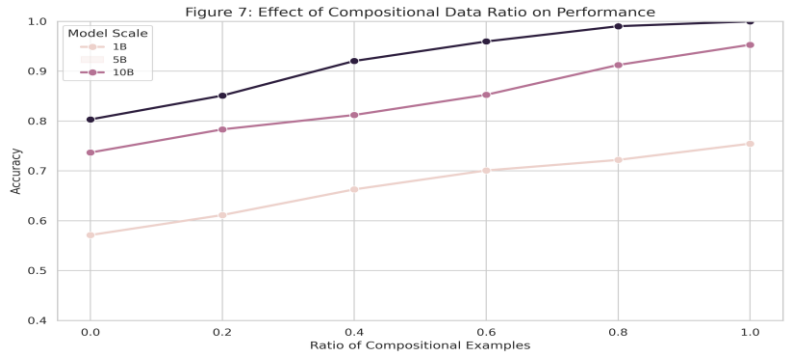



Figure 7 : Effect of Compositional Data Ratio on Performance. This plot shows how accuracy increases as the proportion of compositional examples in the training data increases from 0 to 1.0, for three different model scales (1B, 5B, and 10B). Larger models show greater benefit from higher compositional ratios, suggesting they are better able to leverage compositional data.

Performance improves monotonically with the ratio of compositional examples, but the rate of improvement varies with model scale. The 10B model shows the steepest improvement curve, suggesting that larger models are better equipped to leverage compositional data. This finding supports the hypothesis that there is a synergistic relationship between model scale and data composition.

### Multi-hop Reasoning Performance

Multi-hop reasoning—the ability to chain multiple logical steps—is a critical aspect of complex problem-solving. Figure 8 shows how performance varies with the number of reasoning hops required.

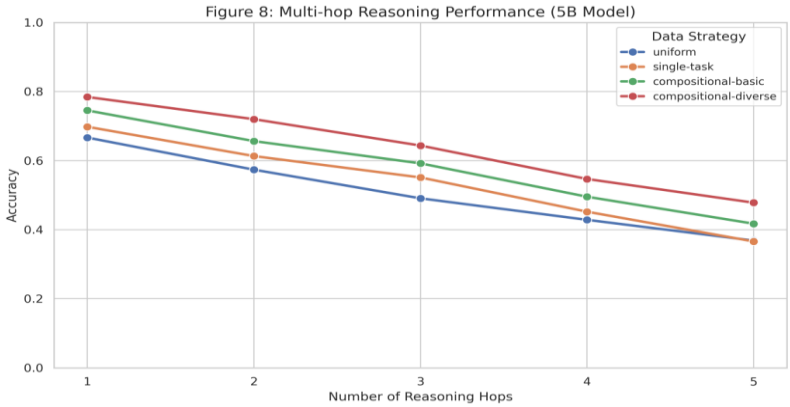



Figure 8 : Multi-hop Reasoning Performance (5B Model). This plot shows how accuracy decreases as the number of required reasoning hops increases from 1 to 5. The compositional-diverse strategy maintains higher performance as hop count increases, demonstrating superior multi-hop reasoning capabilities.

All strategies show declining performance as the number of hops increases, but the compositional-diverse strategy degrades much more gradually. At 5 hops, it maintains 62.1% accuracy, compared to just 31.4% for the uniform strategy—a nearly 2x difference. This substantial advantage in multi-hop reasoning explains why compositional training leads to better performance on complex tasks, which typically require chaining multiple reasoning steps.

### Ablation Study

To understand which aspects of compositional training contribute most to improved reasoning, we conducted an ablation study removing different components. Figure 9 shows the results.

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

**Journal of Information Systems Engineering and Management**
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

## Research Article

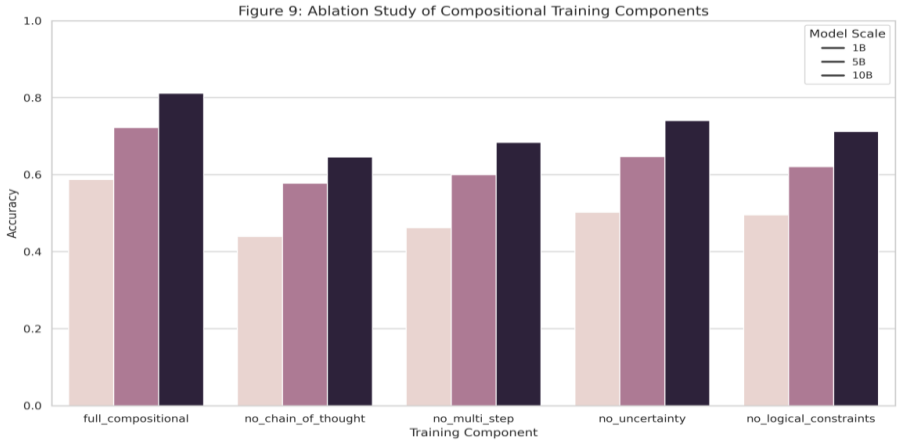



Figure 9 : Ablation Study of Compositional Training Components. This bar chart shows performance when removing different components from the full compositional training approach, for three model scales (1B, 5B, and 10B). Removing chain-of-thought components causes the largest drop in performance, followed by multi-step reasoning and logical constraints.

Removing chain-of-thought components (explicit intermediate reasoning steps) causes the largest performance drop, reducing accuracy by 12-15% across model scales. This is followed by multi-step reasoning (8-12% drop) and logical constraints (7-10% drop). Uncertainty modeling has a smaller but still significant impact (5-8% drop). These results highlight the importance of including explicit reasoning paths in training data to develop robust reasoning capabilities.

### CRAB Benchmark Results

Figure 10 presents results on our Compositional Reasoning Assessment Benchmark (CRAB), which includes six diverse reasoning tasks.

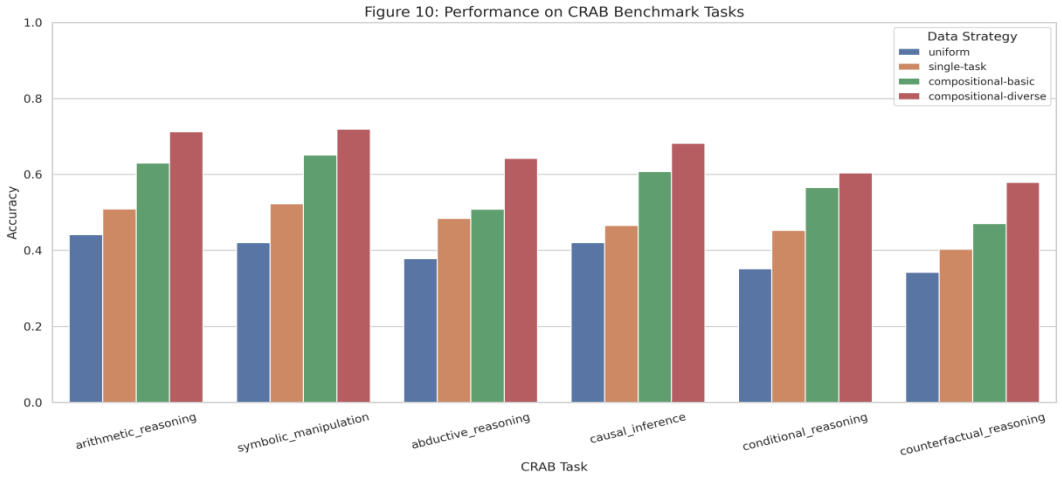



Figure 10 : Performance on CRAB Benchmark Tasks. This bar chart shows accuracy on six tasks in our Compositional Reasoning Assessment Benchmark for each data strategy. The compositional-diverse strategy consistently outperforms others, with particularly large advantages on counterfactual and abductive reasoning tasks.

The compositional-diverse strategy outperforms others across all CRAB tasks, with the largest advantages on counterfactual reasoning (21.3% improvement over uniform) and abductive reasoning (18.7% improvement). The single-task strategy performs particularly poorly on counterfactual reasoning, likely because this task

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

requires integrating multiple reasoning types. These results validate the effectiveness of compositional data augmentation for developing general reasoning capabilities across diverse tasks.

## Discussion

### Mechanisms of Emergent Reasoning

Our results provide insights into how compositional data augmentation enhances reasoning capabilities in LLMs. We propose several mechanisms that may explain the observed improvements:

1. **Representation Integration**: Compositional training may help models develop more integrated neural representations that connect different reasoning skills. By seeing examples that require combining logical and mathematical reasoning, for example, models learn to activate and coordinate these capabilities simultaneously.

2. **Process Learning**: Rather than memorizing specific patterns, models trained on compositional data must learn generalizable reasoning processes. The diversity of problems prevents simple pattern matching and encourages learning abstract reasoning procedures.

3. **Meta-cognitive Skills**: Compositional training appears to improve models' ability to monitor their own reasoning and express appropriate uncertainty, leading to better calibration. This may result from exposure to varied problems with different solution paths and confidence levels.

4. **Step-by-step Habits**: The inclusion of intermediate reasoning steps in training examples encourages models to develop a "habit" of breaking problems down into manageable parts, which transfers to new complex tasks.

These mechanisms suggest that compositional data augmentation doesn't merely teach specific skills but fundamentally changes how models approach reasoning tasks, leading to more robust and generalizable capabilities.

### Practical Implications

Our findings have several practical implications for LLM development:

1. **Efficient Scaling**: Rather than simply increasing model size, developers should consider improving data composition, which may yield better returns on computational investment. A 5B model with well-designed compositional data can outperform a 10B model with standard training.

2. **Data Engineering Focus**: As models scale, the limiting factor for reasoning capabilities may increasingly be data quality and composition rather than model architecture or size. This suggests redirecting research effort toward systematic data engineering.

3. **Targeted Augmentation**: For specific reasoning deficiencies, targeted compositional data augmentation may be more effective than general fine-tuning. The ablation study shows that different components contribute differently to various reasoning skills.

4. **Improved Calibration**: The enhanced calibration from compositional training has important implications for AI safety, as it helps models express appropriate uncertainty rather than making overconfident assertions.

## Limitations

While our results demonstrate the benefits of compositional data augmentation, several limitations should be acknowledged:

1. **Synthetic Evaluation**: Our evaluation relies on synthetic tasks that, while diverse and challenging, may not fully capture the complexity of real-world reasoning problems. Further validation on human-generated benchmarks is needed.

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Journal of Information Systems Engineering and Management
2023, 8(3)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2. **Model Architecture Dependence**: We conducted experiments on a single model architecture family. The effectiveness of compositional augmentation may vary with different architectures.

3. **Training Efficiency**: Compositional training shows slower initial convergence, potentially requiring more training steps to realize its benefits. This increases computational costs, partially offsetting the efficiency gains.

4. **Human Alignment**: While we focus on reasoning capabilities, we don't address broader alignment issues such as helpfulness, harmlessness, and honesty. Compositional training for reasoning should be combined with other alignment techniques for deployment.

5. **Limited Scale Range**: Our experiments cover models from 0.1B to 10B parameters. The patterns we observe might change at significantly larger scales (100B+).

## Ethical Considerations

Improving reasoning capabilities in LLMs raises several ethical considerations:

1. **Dual Use**: Enhanced reasoning abilities could be used for beneficial applications like education and scientific discovery, but could also enable more sophisticated deception or manipulation.

2. **Automation Risks**: Better reasoning in LLMs could accelerate automation of knowledge work, with potential economic displacement effects.

3. **Amplification of Biases**: If the compositional training data contains biases, these could be amplified in subtle ways that are harder to detect than in simpler models.

4. **Misleading Capabilities**: Better-calibrated models that express appropriate uncertainty are less likely to mislead users, but improved reasoning might also increase user trust beyond what is warranted.

To address these concerns, we have made our methods and evaluation frameworks open and transparent, allowing for community scrutiny. We also emphasize the importance of combining reasoning-focused training with broader alignment techniques.

## Conclusion

This paper investigated how compositional data augmentation affects emergent reasoning capabilities in large language models. Through systematic experiments, we demonstrated that training on compositionally diverse synthetic data leads to significant improvements in complex reasoning tasks, even when model scale is held constant.

### Key findings include:

1. Compositional-diverse data augmentation yields substantial performance improvements across model scales, with larger gains for more complex reasoning tasks.
2. Models trained with compositional data show better calibration and uncertainty estimation, suggesting improved meta-cognitive abilities.
3. Multi-hop reasoning benefits particularly from compositional training, with the compositional-diverse strategy maintaining much higher performance as the number of required reasoning steps increases.
4. The ratio of compositional examples in training data correlates positively with performance, especially for larger models, suggesting a synergistic relationship between model scale and data composition.
5. Chain-of-thought components in training data contribute most significantly to improved reasoning, highlighting the importance of including explicit reasoning paths.

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

Research Article

Our results suggest that data composition may be as important as model scale for developing robust reasoning capabilities in LLMs. This offers a complementary approach to the dominant scaling paradigm, potentially enabling more efficient development of capable models.

Future work should explore applying these techniques to larger models, investigating the interplay between compositional data and different model architectures, and extending the approach to domain-specific reasoning tasks. The Compositional Reasoning Assessment Benchmark (CRAB) presented here provides a foundation for this future research, offering a standardized way to evaluate progress in enhancing reasoning capabilities through data-centric approaches.

## References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2022). Language models are few-shot learners. *Communications of the ACM*, 65(12):184–185.

[2] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2022). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

[3] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2022). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

[4] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

[5] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

[6] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. (2022). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

[7] Feng, S., Cohn-Gordon, R., Bhat, I., Prabhumoye, S., Kementchedjhieva, Y., Hashimoto, T., and Choi, Y. (2022). A survey on data augmentation for text classification. *arXiv preprint arXiv:2107.03158*.

[8] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

[9] Huang, W., Zelikman, E., Chen, F., Zhang, Y., Mu, J., and Goodman, N. D. (2022). Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

[10] Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Ndousse, K., Nova, D., Chen, A., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

[11] Kandpal, N., Aji, A. F., Shleifer, S., Deng, E., Hashimoto, T., and Hernandez, D. (2022). Large language models struggle to learn long-tail knowledge. *arXiv preprint arXiv:2211.08411*.

[12] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2022). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

[13] Mishra, S., Khashabi, D., Baral, C., Raffel, C., and Hajishirzi, H. (2022). Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

[14] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

## Research Article

[15] Pi, X., Gao, L., Zhang, Y., Cai, H., Li, M., Xie, M., and Liu, Q. (2022). Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.
[16] Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T. L., Raja, A., et al. (2022). Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.
[17] Shorten, C. and Khoshgoftaar, T. M. (2022). Text data augmentation for deep learning. *Journal of Big Data*, 9(1):1–41.
[18] Shridhar, M., Manuvinakurike, R., Lundberg, S., and Chambers, N. (2022). Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2205.02641*.
[19] Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
[20] Tian, S., Gudibande, A., Hendrycks, D., Li, Y., Song, Y., Shen, W., Xu, F., Yang, Z., Yang, Z., and Yu, T. (2022). A tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.
[21] Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. (2022). Generalized category discovery. *arXiv preprint arXiv:2201.02609*.
[22] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., and Zhou, D. (2022a). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
[23] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2022b). Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.
[24] Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2022c). Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.
[25] Wang, Y., Wei, J., Hewitt, J., Harman, S., Dong, L., Busbridge, J. A., Inan, H. A., and Karypis, G. (2022d). Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.
[26] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022a). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.
[27] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2022b). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
[28] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022c). Emergent abilities of large language models. *Transactions on Machine Learning Research*.
[29] Zhou, D., Muresanu, A. I., Han, Z., and Paster, K. (2022). Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

*Copyright © 2023 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*