Local LLMs: True Reasoners or Just Memorizers?

Bickston Laenger (bickston@gatech.edu), Araceli Rodriguez Vallejo (avallejo32@gatech.edu), Aditya Raghavan (araghavan68@gatech.edu) , Juan Macias Romero (jromero74@gatech.edu) , Alejandro Danies-Lopez (adanieslopez3@gatech.edu)

1 Introduction

Large language models have seen dramatic growth in size and capability over the past few years, moving from hundreds of millions to hundreds of billions of parameters. In their design, there is often a trade-off between model performance and the computational cost required to train and serve them. Distilled models try to get the best of both worlds by having significantly fewer parameters while keeping a large amount of the knowledge. We wish to compare these models to explore how the current best distilled models perform. By comparing models, under the scope of reasoning tasks and across parameter sizes and architectures, we can better understand how scale and design choices translate into real world performance on high-difficulty scenarios.

Knowledge distillation is a process by which a large "teacher" model transfers its learned behaviors and representations to a smaller "student" model. Originally formalized by Hinton et al. (Hinton et al., 2015), the student model is trained to match the teacher's softened output distributions, allowing it to approximate the teacher's performance with fewer parameters and lower latency. In this project, each open-source model (Deepseek-R1, Gemma 3, and LLaMA 3) is evaluated in two of its distilled smaller-parameter incarnations, comparing 1–1.5 B-parameter students against larger 4–8 B-parameter fellow students.

2 Literature Review

Large language models (LLMs) have grown dramatically in scale, evolving from GPT-2's 100 million parameters to models exceeding 500 billion, such as Google's PaLM (Chowdhery et al., 2022). Yet, as models get larger, the question arises: do they actually get better at reasoning? Or are we merely overfitting language to look intelligent? While scaling has undeniably improved performance on traditional benchmarks, models still underperform on tasks that demand structured reasoning and abstract symbolic manipulation.

To test this, the BIG-Bench project (Srivastava and et al., 2022) introduced over 200 tasks designed to test a wide spectrum of language understanding and reasoning capabilities. Models ranging from GPT-3 to PaLM were evaluated in few-shot settings without fine-tuning, and performance was manually rated against human benchmarks. The follow-up benchmark, BIG-Bench Hard (Suzgun et al., 2022), distilled this larger set into 23 tasks where even the best-performing models consistently underachieved. These tasks include mostly problems that require not just pattern recognition but genuine structural reasoning.

The question of whether smaller models can faithfully replicate their larger counterparts is central to the study of knowledge distillation. Originally proposed by (Hinton et al., 2015), distillation has since evolved into a sophisticated strategy for compressing models while preserving performance. TinyBERT (Jiao et al., 2020), DistilBERT (Sanh et al., 2020), and other variants have shown that distilled models can retain high performance on general NLP benchmarks. However, when it comes to structured reasoning tasks, the results are more mixed. (Roberts et al., 2025) find that smaller models often fail to retain logical consistency and hierarchical understanding, especially in symbolic domains like BBH. This suggests that while distillation can preserve surface-level competence, deeper reasoning skills may require both scale and architectural nuance.

3 Methodology

We began by downloading models from Ollama's supported registry, which allowed us to run them locally. Initially, we aimed to test larger models like Gemma-12B and Deepseek-R1-14B, but quickly ran into hardware limitations like frequent crashes and execution failures. We also considered including proprietary models such as GPT-4 and Claude via their APIs as external baselines. However, strict token limits, rate caps, and high pricing prevented us from incorporating them into our analysis. As a result, we reduced our maximum model size to 7–8 billion parameters. Specifically, we used:

- **Deepseek-R1**: 1.5B and 8B
- **Gemma 3**: 1B and 4B
- **LLaMA 3.2**: 1B and **LLaMA 3.1**: 8B

Our initial experiments involved manually prompting a few hundred examples across simpler tasks like SAT-level verbal and math questions, along with basic arithmetic and logical reasoning. Across all six models, accuracy on these tasks was consistently high, ranging from 85% to 95%, with minimal performance difference between smaller and larger models. This prompted a shift toward more challenging evaluation domains.

To better differentiate the capabilities of the models and probe their reasoning depth, we selected 10 tasks from the BIG-Bench Hard (BBH) suite (Srivastava and et al., 2022). BBH tasks are specifically designed to test the limits of reasoning, logic, and language comprehension. Due to computational constraints and time limitations, we opted not to run all 23 BBH tasks, instead selecting a representative and diverse subset spanning logical deduction, multi-step reasoning, and deceptive prompt scenarios. Python scripts were developed to dispatch model queries in parallel batches and log responses in a structured format for later analysis, with each student taking a subset of tasks to run on their own.

The analysis consisted of manual annotation to determine the correctness of each response. While some tasks allowed for programmatic accuracy checking (e.g., arithmetic problems), most required human evaluation due to the nuanced nature of reasoning and multi-part answers.

Each task's result set was analyzed to compute:

- Accuracy (correct responses / total responses)
- Performance across model scale

These annotations allowed us to derive quantitative accuracy scores to see how each model handled complex reasoning under reduced parameter budgets.

3.1 Explanation of each task

Task	Description
Riddles	Predict the correct answer to a lateral thinking problem.
Translation Errors	Given a source sentence written in German and its translation in English, determine the type of translation error.
Dyck Language Correction	Predict the sequence of the closing parentheses of a Dyck-4 word without its last few closing parentheses.
Object Counting	Given a collection of possessions that a person has along with their quantities, determine the number of a certain object/item class.
Formal Fallacies	Given a context involving a set of statements, determine whether an argument can be logically deduced from the provided context.
Web of Lies	Evaluate the truth value of a random Boolean function expressed as a natural-language word problem.
Penguins in a Table	Given a unique table of penguins, answer a question about the attributes of the penguins (such as age, weight, etc.).
Tracking Shuffled Objects	Given the initial positions of a set of objects and a series of pairwise swaps applied to them, determine their final positions.
Adjective Order	Given two English-language sentences, determine the one with the correct adjective order.
Multi-step Arithmetic	Solve multi-step equations involving basic arithmetic operations (addition, subtraction, multiplication, and division).

Table 1: Task Descriptions

4 Model Comparisons

4.1 Riddles

Gemma performed the best at this task with LLaMA behind and Deepseek bringing up the far rear. It does not necessarily seem as though Gemma and LLaMA had more of a riddle base in their training data, but instead that Deepdeek often reasoned itself to the correct answer and then continued reasoning itself out of the correct answer. When given a common riddle but with major details changed, most models still gave the answer to the common riddle, suggesting that the models are answering straight from training data. Further, even telling them the common answer was incorrect in the prompt didn't make them change their answer, further cementing the memorization likelihood.

4.2 Translation Errors

Deepseek performed the best in this task and below it came both LLaMA and Gemma. Deepseek was much better at working through the translations piece by piece when reasoning to determine which part sounded the most incorrect. Gemma and LLaMA on the other hand were not able to reason and instead often had one or two answers (out of a possible seven) that the models answered the majority of the time. This ended with the smaller models answering with the same accuracy as random guessing.

4.3 Dyck Language Correction

Gemma truly outperformed in this task, as its 1B version scored higher than any other non-Gemma model. This suggests that Gemma either had better training data than LLaMA or is generally better at fixing Dyck language strings. This was also the only task in Which LLaMA 1B ended up getting zero correct. A curious pattern that emerged for both LLaMA and Deepseek was that for a substantial number of answers the entire output was correct except the second to last parenthesis, which they excluded. This happened across a wide variety of input sizes, suggesting they may have had a similar error in training.

4.4 Object Counting

In this task, common errors across all models included excluding certain objects from the prompt and basic arithmetic mistakes during counting. Approximately 70% of the errors by Gemma models stemmed from minor math issues despite correct object identification and parsing, showing their strong comprehension abilities. Though LLaMA 3.1 (8B) appears to perform better than 3.2, its errors were more erratic, ranging from misclassifications to flawed assumptions and arithmetic mistakes, suggesting a less consistent reasoning pattern compared to LLaMA 3.2. The Deepseek models exhibited distinctive issues such as self-contradictory reasoning (where the model initially arrived at the correct answer but reversed it upon further reasoning) and overly verbose outputs that exceeded length limits without producing

Benchmark	Avg. Human	Gemma1B	Gemma4B	LLaMA1B	LLaMA8B	Deepseek1.5B	Deepseek8B
True/False Questions							
Web of Lies	0.813	0.500	**0.980**	0.300	0.440	0.520	0.640
Formal Fallacies	**0.908**	0.620	0.469	0.553	0.480	0.744	0.839
Multiple Choice Questions							
Penguins in a Table	0.780	0.300	0.820	0.160	0.400	0.600	**0.920**
Tracking Shuffled Objects	0.647	0.380	0.260	0.280	0.180	0.480	**0.940**
Adjective Order	0.747	0.780	**0.880**	0.560	0.540	0.480	0.640
Translation Errors	0.367	0.200	0.340	0.120	0.420	0.220	**0.600**
Open-ended Questions							
Object Counting	0.240	0.608	**0.736**	0.400	0.460	0.510	0.676
Dyck Language Correction	0.478	0.240	**0.500**	0.000	0.100	0.040	0.220
Riddles	-	0.479	**0.833**	0.375	0.708	0.083	0.167
Multi-step Arithmetic	0.097	0.460	0.720	0.000	0.000	0.740	**0.880**

Table 2: Accuracy of average human raters and different model families and sizes across reasoning benchmarks.

a final result. Roughly 15% of errors stemmed from incorrect assumptions, such as claiming that snakes aren't animals or similar mess ups.

4.5 Formal Fallacies

In this task, no model consistently outperformed the others, and overall accuracy remained relatively low across the board. Deepseek models frequently attempted to reason through arguments multiple times, often resulting in self-contradiction or confusion. In roughly 30% of prompts, the response limit was reached with no final answer. However, when Deepseek did produce a response, it was correct approximately 80% of the time, indicating strong potential when not hindered by verbosity. Gemma and LLaMA models showed no clear correlation between parameter size and accuracy, and their outputs varied unpredictably across different prompts. While some promising reasoning patterns were observed, the overall inconsistencies and low average performance suggest a need for a larger and more diverse prompt set, as well as more targeted fine-tuning focused on formal logic and inference skills

4.6 Web of Lies

Gemma4B was dominant for this task, achieving almost perfect accuracy (the highest of any model at any task) and beating the average of human raters from (Suzgun et al., 2022). Gemma1B, however, had performance equal to random chance, showing that model size is important for this task. Deepseek models were moderately competent, with both outperforming Gemma1B but not human raters. As in other tasks, a common failure mode for Deepseek involved it questioning its own reasoning and going through loops of rephrasings and verification. Both LLaMA models, on the other hand, performed worse than random chance. These were the most likely to say there wasn't enough information to answer the question, or to try to map the questions to another problem with no correct answer, but in most cases an incorrect answer was given.

4.7 Penguins in a Table

For all three model families, performance in this task scaled directly with parameter count. Deepseek performed the best, with the 8 billion parameter model having 92% accuracy. Though Gemma had slightly worse performance, with 82% accuracy for the 4B model, its gains from scale were particularly noteworthy, considering the 1B model's accuracy was only 30%—an increase of over 50 percentage points when quadrupling parameter count. The other models saw much more modests improvements even though the parameter count increase was larger, suggesting Gemma models gain the most from scale when used for this type of attribute retrieval and analysis task.

4.8 Tracking Shuffled Objects

This task is particularly demanding because it requires precise, multi-step reasoning and symbolic state tracking. While most models struggle, Deepseek8B stands out with a 94% accuracy. In contrast, Deepseek1.5B performs at 48%, but what's notable is how it fails: rather than confidently guessing, it often expresses uncertainty, and in 38% of its incorrect responses, the correct answer is present in the reasoning but not selected — indicating that it understands the logic but struggles with final decision-making.

LLaMA 1B and Gemma models attempt step-by-step reasoning, but fail for different reasons. LLaMA1B often starts confidently but makes mistakes in the early swaps that derail the entire chain. Gemma1B follows a similar structure but tends to make more subtle tracking errors or mix up positional references, such as left and right. In contrast, LLaMA3.2 8B rarely attempts structured reasoning at all, often jumping straight to a final answer with little explanation.

4.9 Adjective Order

Gemma models perform noticeably better than both LLaMA and Deepseek, with Gemma4B achieving the highest accuracy at 86.3%. Both Deepseek and Gemma often reference the adjective ordering rule explicitly and justify their choices by categorising adjectives. However, results LLaMA models tend to choose what "sounds right," often using phrases like "this feels more natural" without providing any rule-based justification. It only applies proper adjective ordering logic in a small fraction of its responses.

4.10 Multi-step Arithmetic

When evaluating DeepSeek's responses, we noticed that in many cases, Deepseek's final output didn't seem to be conditioned by its reasoning, intermediate output. In fact, in most cases, the model directly predicted a response value to the query. We decided to analyze both as "thinking output" and "final response".

In some rare cases, the model outputted a short description of the steps to be taken, without directly trying to reach a response. This only occurred in 2 of the 50 responses. Excluding these, we found that, in 41% of the cases where the 1.5b parameter version was wrong in its final predicition, it had actually reached the correct answer while thinking. This behavior is less frequent in its larger, 7b counterpart - 16%. This suggests that larger models also improve in their ability to extract knowledge from its reasoning output.

5 Model Family Comparisons

Overall, the Gemma models showed the best performance for their respective sizes. It is particularly noteworthy that the 4 billion parameter model outperformed the other large models, even though they had twice its parameter count. In 5, the line corresponding to Gemma has the largest slope, indicating that out of the three models, it gains the most from scale. The web of lies and penguins benchmarks showed the greatest improvement with increased model size for Gemma. These tasks involve understanding relationships between newly presented items of information, suggesting that this is the cognitive skill that most benefits from more computation at this scale out of the ones we tested. The shuffle objects and formal fallacies, however, showed decreased performance with increased parameter count, highlighting the fact that, due to the complex nature of these models, the relationship between model size and performance isn't always linear.

It is also worth highlighting that Gemma models had lower variability in accuracy for open-ended questions than other models, and had higher performance for this type of questions than for multiple choice ones. It seems like the more constrained solution spaces of multiple choice questions don't provide much assistance to Gemma models, unlike humans and other models.

The LLaMA models, on the other hand, consistently had the worst performance, with neither one achieving the best performance on any given metric and the average performance of the large LLaMA model being worse than that of the smaller models of the other families. LLaMA models had an accuracy of zero for multi-step arithmetic problems, and the smaller one had the same score for the language correction benchmark; even at true/false or multiple choice problems, their performance was often worse than chance. Performance for adjective order and (as with Gemma) shuffled objects and formal fallacies decreased with scale, often due to the models resorting to guessing instead of reasoning.

Deepseek models uniformly gained from increase

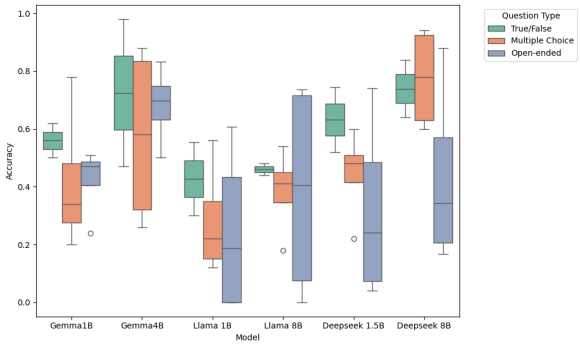

Figure 1: Model accuracy by type of question

model size across tasks, likely due to the larger model's improved reasoning capabilities. These models did outperform Gemma at multiple choice problems, but had much more variable accuracy for open-ended ones and a worse average, suggesting that Deepseek models might be more appropriate for tasks with more constrained response sets.

6 Conclusion

The results from our experiments demonstrated that scale generally (but not universally) improves performance, albeit with diminishing returns on more complex reasoning. Larger models almost always outperform their smaller counterparts on straightforward retrieval tasks.

Gemma 3 shows the strongest scale-up benefit on relational and logical tasks, as well as the best open-ended problem performance. The jump from Gemma 1B to Gemma 4B yields the largest absolute gains on web of lies and penguins in a table, suggesting that its distilled training preserves higher-order relational understanding more effectively than other families.

Deepseek-R1 scales consistently but sometimes overthinks. Across all tasks, the 8B variant outperforms the 1.5 B one, demonstrating the benefits of additional computational capacity. We still observed frequent loops of self-contradiction in its reasoning traces, particularly on verbose tasks, a spot for further improvement.

LLaMA's scale-up gains are task-dependent and occasionally non-monotonic. While LLaMA 8B matches or exceeds the 1B model on most benchmarks, it paradoxically underperforms on tasks where the smaller model's structured-reasoning attempts yield better results. This suggests that beyond a certain size, LLaMA may default to heuristic guessing rather than systematic reasoning.

These findings show the nuanced trade-offs in distilling and scaling language models for reasoning tasks, and that true structural reasoning remains a frontier for smaller-scale models. Ultimately, bridging the gap between parameter efficiency and strong reasoning will require both architectural innovation and more targeted distillation methodologies.

References

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. *Preprint*, arXiv:1909.10351.

Nicholas Roberts, Niladri Chatterji, Sharan Narang, Mike Lewis, and Dieuwke Hupkes. 2025. Compute optimal scaling of skills: Knowledge vs reasoning. *Preprint*, arXiv:2503.10061.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Aarohi Srivastava and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

A Appendix: Graphs on Model Accuracy

This section contains additional graphs comparing model accuracy across different models and parameter sizes.

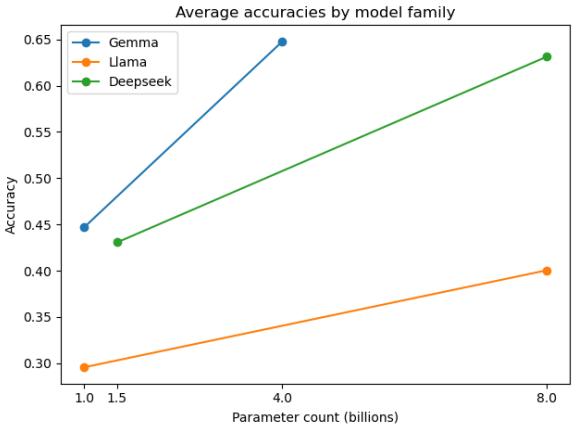

Figure 5: Model accuracy by family and size

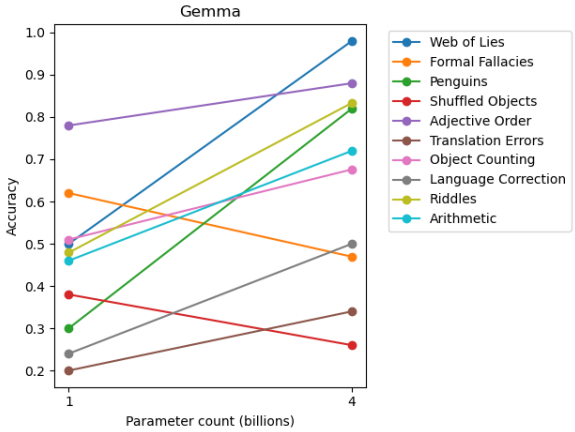

Figure 2: Gemma accuracy vs. parameter count across benchmarks

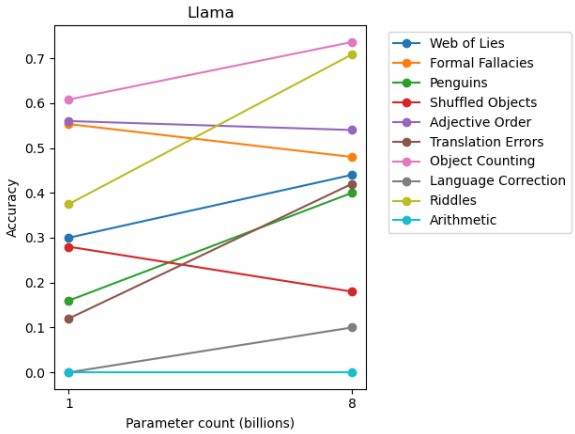

Figure 3: LLaMA accuracy vs. parameter count across benchmarks

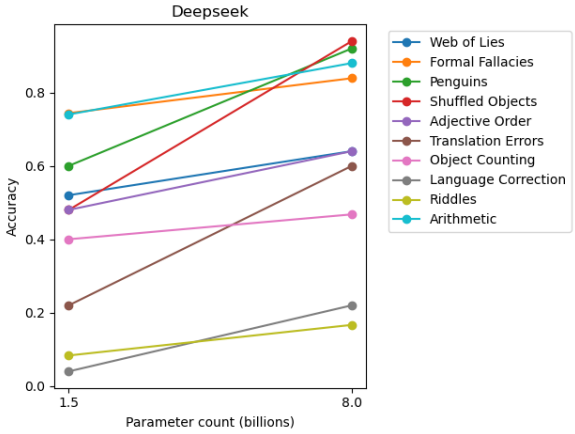

Figure 4: Deepseek accuracy vs. parameter count across benchmarks