EPiC Series in Technology

Volume 3, 2026, Pages 36–49

Proceedings of AI for Accelerated Research Symposium

OmniScience: A Domain-Specialized LLM for Scientific Reasoning and Discovery

Vignesh Prabhakar[1]*, Md Amirul Islam[1]*, Adam Atanas[1]*, Yao-Ting Wang[1], Joah Han[1], Aastha Jhunjhunwala[2], Rucha Apte[2], Robert Clark[2], Kang Xu[1], Zihan Wang[2], and Kai Liu[1†]

[1] SES AI
[2] NVIDIA

Abstract

Large Language Models (LLMs) have demonstrated remarkable potential in advancing scientific knowledge and addressing complex challenges. In this work, we introduce OmniScience, a specialized large reasoning model for general science, developed through three key components: (1) *domain adaptive pretraining* on a carefully curated corpus of scientific literature, (2) *instruction tuning* on a specialized dataset to guide the model in following domain-specific tasks, and (3) *reasoning-based knowledge distillation* through fine-tuning to significantly enhance its ability to generate contextually relevant and logically sound responses. We demonstrate the versatility of OmniScience by developing a battery agent that efficiently ranks molecules as potential electrolyte solvents or additives. Comprehensive evaluations reveal that OmniScience is competitive with state-of-the-art large reasoning models on the GPQA Diamond and domain-specific battery benchmarks, while outperforming all public reasoning and non-reasoning models with similar parameter counts. We further demonstrate via ablation experiments that domain adaptive pretraining and reasoning-based knowledge distillation are critical to attain our performance levels across benchmarks.

1 Introduction

Large Language Models (LLMs) (OpenAI, 2022; Dubey et al., 2024; Team et al., 2023; Anthropic, 2023; Guo et al., 2025; OpenAI, 2024; xAI, 2025; OpenAI, 2025) have demonstrated widespread success across scientific fields (Bolton et al., 2024; Taylor et al., 2022; Feng et al., 2024; Chithrananda et al., 2020; Zhang et al., 2025, 2024; Tang et al., 2025), from summarizing research papers to generating hypotheses. In battery research, for example, a science-focused LLM could rapidly screen molecular datasets, explore chemical spaces, and identify promising electrolyte candidates, streamlining discovery workflows.

*Joint first authors
†Corresponding author: kai.liu.ai4science@gmail.com

However, developing a highly specialized LLM for general science poses unique challenges. General-purpose foundation models (Dubey et al., 2024; OpenAI, 2022; Team et al., 2023; Anthropic, 2023; Guo et al., 2025) often lack the domain-specific vocabulary and contextual understanding needed to tackle complex scientific topics, such as molecular structures, electrochemical properties, experimental data, and gene expression patterns. While domain-specific language models like ChipNeMo (Liu et al., 2023) for chip design and BloombergGPT (Wu et al., 2023) for financial data have shown the benefits of specialized adaptation, similar efforts in general science remain limited (Taylor et al., 2022; Feng et al., 2024; Zhang et al., 2024; Frey et al., 2022; Li & Jiang, 2021; Yunusoglu et al., 2025; Zhang et al., 2025; Tang et al., 2025). This gap highlights the need for a dedicated general science LLM to address the specific demands of the field.

Training an LLM from scratch is computationally prohibitive. Domain Adaptive Pre-Training (DAPT) (Gururangan et al., 2020) offers a practical alternative by further training a foundation model on a targeted corpus (Shen et al., 2024; dos Santos Junior et al., 2024; Liu et al., 2023), achieving significant domain-specific improvements at a fraction of the cost while retaining general language understanding.

In this work, we present OmniScience (see Fig. 1), an adaptation of the LLaMA 3.1 70B model (Dubey et al., 2024) specifically tailored for science exploration. Our approach leverages DAPT on a carefully curated dataset that includes peer-reviewed articles, arXiv papers, journals, and textbooks covering general science and electrochemistry. To ensure high-quality input, we develop a robust data-processing pipeline that cleans and organizes the text effectively. We then perform DAPT on this domain-specific corpus, enabling the model to acquire a deep understanding of scientific language and concepts.

Following standard practice (OpenAI, 2022; Dubey et al., 2024), we refine our base OmniScience through instruction tuning on both science-specific and general chat instructions, yielding OmniScience Chat.

Recent work (Team et al., 2025; Guo et al., 2025; OpenAI, 2024; xAI, 2025; OpenAI, 2025) has enabled models to engage in extended internal *thinking*—iterative reasoning processes such as hypothesis generation, chain-of-thought analysis, and self-correction—significantly boosting *reasoning* capabilities. To leverage this, we add a knowledge distillation stage on top of OmniScience Chat using the s1K-1.1 dataset (Muennighoff et al., 2025), originally derived from DeepSeek-R1 (Guo et al., 2025) reasoning traces. This reasoning-based distillation bridges the gap between basic instruction alignment and advanced inferential performance, enabling our

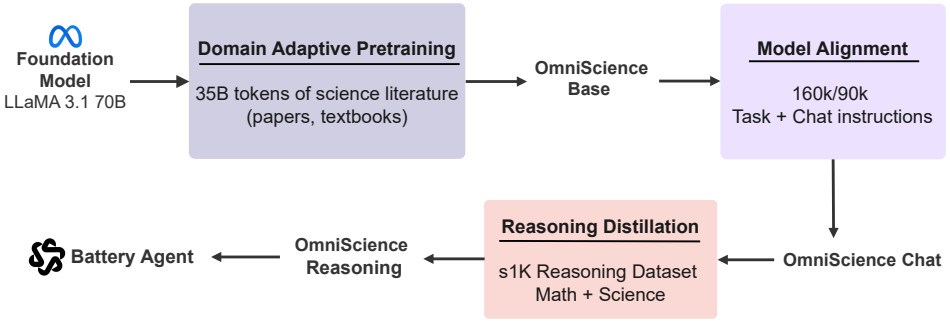

Figure 1: Illustration of our OmniScience training pipeline. We begin with a LLaMA 3.1 70B foundation model, apply domain adaptive pretraining to obtain the OmniScience base model, and then perform model alignment and reasoning-based knowledge distillation to produce the final OmniScience Reasoning model.

model to achieve competitive results on science reasoning tasks. OmniScience achieves 0.72 on GPQA-Diamond, state-of-the-art among similarly-sized models.

We demonstrate OmniScience's adaptability by applying it to battery research—ranking molecules and explaining their suitability as electrolyte solvents or additives, bridging the gap between general-purpose LLMs and the specific needs of scientific discovery.

2 Related work

Recent advancements in large language models (LLMs) have accelerated growing interest in developing domain-specific LLMs, driven by the availability of vast public and proprietary datasets. Early efforts include models, such as BloombergGPT (Wu et al., 2023) for finance, BioMedLLM (Bolton et al., 2024) for biomedical applications, and Galactica (Taylor et al., 2022) for general scientific research. These models were trained from scratch on raw, domain-specific datasets, and although effective, they require billions of tokens and substantial computational resources to achieve acceptable performance. In parallel, domain-specific models in the chemistry space have emerged, including ChemBERTa (Chithrananda et al., 2020), MolBERT (Li & Jiang, 2021), ChemGPT (Frey et al., 2022), and ChemLLM (Zhang et al., 2024). These models leveraged transformer architectures and specialized pretraining to capture chemical knowledge, enabling tasks like property prediction and reaction modeling, though they too often require extensive computational budgets. Despite these promising advances, relatively little effort has been devoted to developing a science-focused LLM that can be easily adapted to specific scientific tasks. Our work seeks to address this gap by creating a model that not only excels in general scientific reasoning but can also be tailored to specialized applications, such as battery research.

While our work focuses on model weights, retrieval-augmented approaches (Lewis et al., 2020; Borgeaud et al., 2021; Izacard et al., 2022; Liu et al., 2023) can further boost performance. We also demonstrate our model in a full agentic setup with RAG for molecular ranking.

3 OmniScience: Domain Adaptive Reasoning Model

We describe our three-step training pipeline for building OmniScience. First, we perform continuous pretraining (Sec. 3.1) on a large corpus of scientific literature to provide the model a solid foundation in domain-specific language and concepts. Next, we perform supervised fine-tuning (Sec. 3.2) using a carefully curated instruction dataset, which aligns the model's responses with high-quality, contextually relevant information. Finally, we apply an extra fine-tuning step (Sec. 3.3) on a high-quality reasoning dataset to further improve the model's reasoning ability to handle complex scientific tasks.

We deduplicated and excluded all evaluation datasets used for benchmarking from DAPT/SFT/distillation corpora. Consequently, all reported battery results are zero-shot.

3.1 Domain Adaptive Pretraining

Pretraining Data. Our 35B-token corpus spans 3.6M peer-reviewed articles, arXiv/ChemRxiv preprints, PubChem, Semantic Scholar (Lo et al., 2020), and academic textbooks, ensuring broad scientific domain coverage (details in Appendix A.2).

Data Processing. PDFs are converted to text (Unstructured.IO, 2022), deduplicated via MinHash (Broder, 1997) and LSH (Gionis et al., 1999), filtered with NeMo Curator (Kuchaiev et al., 2019), and tokenized using LLaMA 3.1 70B tokenizer (Dubey et al., 2024).

Model Architecture. We initialize from LLaMA 3.1 70B base (Dubey et al., 2024), training with NVIDIA NeMo (Guillaume & Rougemont, 2006) using tensor parallelism (Shoeybi et al., 2019), flash attention (Dao et al., 2022), and context parallelism for 8K-token sequences, yielding **OmniScience base**.

Training Details. Learning rate: 1×10^{-5}; optimizer: AdamW (Kingma & Ba, 2014); weight decay: 0.0001; global batch: 64 (524K tokens/step); 66K steps (1 epoch on 35B tokens); bfloat16 precision; 6 days on 128 H100 GPUs. Training curves in Appendix A.1.

3.2 Model Alignment with Supervised Fine-Tuning (SFT)

SFT Data. We sample 50K papers from our DAPT corpus and use GPT-4o-mini to generate 200K instruction samples (Q/A, summarization, reading comprehension, MCQ). We retain 160K for training and reserve 40K for held-out evaluation. Combined with 90K Daring Anteater chat samples (Wang et al., 2024), our SFT dataset comprises 250K samples.

SFT Training. Using NeMo with learning rate 1×10^{-6}, batch size 64, we fine-tune OmniScience base for 1400 steps (32 hours on 128 H100 GPUs), yielding **OmniScience Chat**. Training curves in Appendix A.1.

3.3 Reasoning-based Knowledge Distillation

s1K Dataset. The s1K-1.1 dataset (Muennighoff et al., 2025) is a curated 1K-sample subset from 59K DeepSeek-R1 traces, refined through quality filtering, difficulty selection (questions both Qwen2.5-7B/32B fail), diversity clustering, and decontamination against GPQA Diamond and other benchmarks.

Fine-tuning. We fine-tune OmniScience Chat on s1K-1.1 using 16K context (fits 94% of reasoning traces), batch size 2, learning rate 1×10^{-5} (128-step warmup, cosine decay), AdamW (β_1=0.9, β_2=0.95, weight decay 0.0001), for 5 epochs (2500 steps, 12 hours on 64 H100 GPUs), yielding **OmniScience Reasoning**. This distillation transfers DeepSeek-R1's reasoning capacity to our compact model.

4 Experiments

In this section, we present a comprehensive evaluation of our model's performance on both public (Sec. 4.1) and domain-specific (Sec. 4.2) benchmarks. We benchmark against state-of-the-art models using our OmniScience Reasoning model and rigorously assess its capabilities across diverse tasks and datasets. Additionally, we perform ablation studies (Sec. 4.3) to demonstrate the necessity of our continuous pretraining and reasoning alignment steps. Finally, we show the adaptability of our OmniScience to solve battery related tasks in Sec. 4.4.

Evaluation policy: Unless explicitly stated, *all evaluations are non-RAG* and use a single unified prompt template, decoding policy and context length across models for apples-to-apples comparisons.

4.1 Results on Public Benchmarks

In Table 1, we compare our results with recent state-of-the-art reasoning and non-reasoning models on GPQA Diamond benchmark. GPQA Diamond (Rein et al., 2024) is particularly relevant for our science-focused LLM, as our post-training was specifically tailored to domain-specific scientific tasks. Our OmniScience Reasoning model achieves a score of 0.720, narrowly

surpassing DeepSeek-R1 (0.715), despite having only 10% of DeepSeek-R1's parameters and not requiring pretraining from scratch. Notably, our model remains competitive with much larger reasoning models, such as Claude 3.7 sonnet (0.782), Grok 3 (0.802), and GPT-o1 (0.757). Our model also outperforms most non-reasoning models, even those with huge sets of parameters such as Llama-3.1-405B, DeepSeek-v3, and GPT-4.5. We believe that further scaling of our architecture will yield even greater performance gains.

Fig. 2 compares models in the 10–100B range. OmniScience surpasses all baselines including DeepSeek-R1 Distill LLaMA 70B (0.720 vs. 0.652), despite starting from the base (not instruct) LLaMA-3.1-70B, underscoring the effectiveness of DAPT combined with reasoning distillation.

Methods	Params	GPQA
GPT-o1 (OpenAI, 2024)	-	0.757
o3-mini-low (OpenAI, 2025)	-	0.706
DeepSeek-R1 (Guo et al., 2025)	685B	0.715
Claude 3.7 (thinking)	240B	0.782
Gemini 2.0 (Team et al., 2023)	-	0.742
Grok 3 (thinking)	2.7T	0.802
s1 (Muennighoff et al., 2025)	32B	0.636
GPT-4o (OpenAI, 2024)	-	0.499
LLaMA 3.1 (Dubey et al., 2024)	405B	0.507
DeepSeek-v3 (Guo et al., 2025)	685B	0.591
GPT-4.5 (OpenAI, 2022)	-	0.714
Grok-3 (no thinking)	-	0.754
Claude-3.7 (no thinking)	240B	0.680
OmniScience Reasoning	**70B**	**0.720**

Table 1: Performance comparison between different SOTA reasoning models on the GPQA Diamond benchmark.[1]

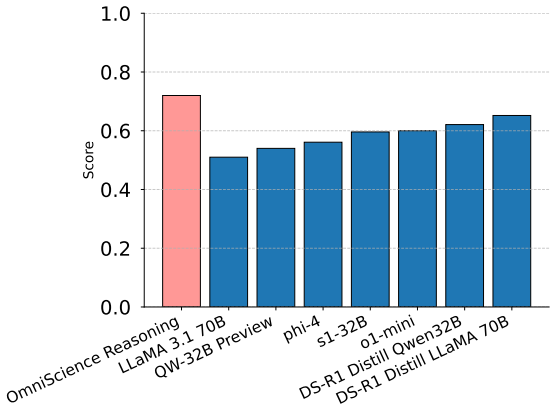

Figure 2: Comparison of GPQA Diamond scores with top 10-100B parameter models. Our model outperforms all the baselines including DeepSeek-R1 distill variants.

Benchmark breadth Beyond GPQA-Diamond, we evaluated OmniScience on two public science benchmarks. *Mol-Instructions* (Fang et al. (2023)) for biomolecular instruction understanding and *ScienceAgentBench* (Chen et al. (2024)) for agentic scientific reasoning show that our 70B model generalizes beyond electrochemistry/materials: on Mol-Instructions we match or exceed open-weight baselines (vs. DeepSeek-R1 and open o1 variants) across entity/disease/protein subtasks, and on ScienceAgentBench we substantially outperform Llama 3.1 70B while approaching stronger proprietary models. Full per-subtask results have been shown in Appendix A.4.

4.2 Results on Battery Domain-Specific Benchmarks

We evaluate on an SFT held-out test set (Sec. 3.2) and an internally curated battery reasoning benchmark, comparing against GPT-o1, LLaMA 3.1 70B, Claude 3.7 Sonnet, and Gemini. As shown in Table 2, OmniScience Reasoning outperforms all baselines except GPT-o1 (which was used to generate SFT data), achieving competitive results despite significantly fewer parameters. Notably, OmniScience Reasoning matches GPT-o1 on Battery Q/A (96%) and Reading Comprehension (90%).

[1]**Note:** Scores reflect our unified prompt + context + non-RAG policy; vendor-reported numbers may differ due to alternate templates, evaluation scripts or retrieval settings.

Model	Battery Q/A	Battery MCQ	Battery RC	Battery Summ	Battery Reasoning
LLaMA 3.1 70B	71%	67%	78%	75%	66%
Claude 3.7 Sonnet	94%	86%	89%	86%	80%
Gemini Flash Thinking	92%	85%	88%	82%	79%
GPT-o1	96%	92%	90%	88%	84%
OmniScience Chat	93%	79%	84%	79%	73%
OmniScience Reasoning	**96%**	**89%**	**90%**	**86%**	**82%**

Table 2: Performance comparison on battery-specific tasks, including Q/A, MCQ, Reading Comprehension, Summarization, and Reasoning.

Methods	GPQA
LLaMA 70B Distillation	0.58
CPT Distillation	0.72
Chat Distillation	**0.72**

Table 3: Performance comparison of model variants on GPQA Diamond benchmark. Results highlight the necessity of continuous pretraining for scientific reasoning and the supplementary benefits of SFT.

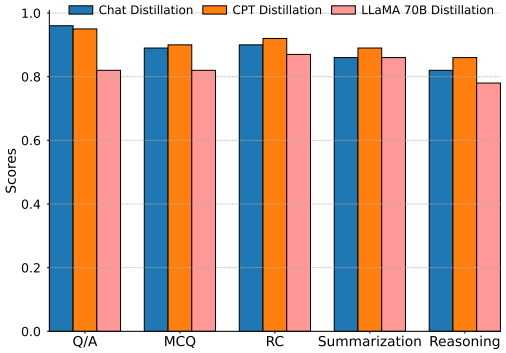

Figure 3: Performance comparison of model variants on battery benchmarks.

4.3 Ablation Studies

To evaluate the impact of domain adaptive pretraining and supervised fine-tuning on our model's performance, we isolate the reasoning distillation-based fine-tuning step. As discussed in Sec. 3.3, we apply this distillation step to our OmniScience Chat model to derive the OmniScience Reasoning model. As an alternative approach, one could directly distill the base LLaMA 3.1 70B model on the s1K dataset, bypassing both domain adaptive pretraining and supervised fine-tuning. To assess these strategies, we compare three model variants: **Chat Distillation** (distillation on the OmniScience Chat model, which produces the OmniScience Reasoning model), **CPT Distillation** (distillation on the OmniScience base model), and **LLaMA 70B Distillation** (direct distillation on the LLaMA 3.1 70B base model). The results in Table 3 show that both the CPT and Chat Distillation models outperform the LLaMA 70B Distillation variant on the GPQA Diamond benchmark. Notably, on tasks requiring complex scientific reasoning, such as GPQA Diamond, the CPT and Chat Distillation models achieve a score of 0.72 compared to 0.58 for the LLaMA 70B Distillation model, highlighting the importance of continuous pretraining and supervised fine tuning for effective domain specific reasoning.

In addition to the public GPQA Diamond benchmark, we evaluate the models on battery specific tasks (see Fig. 3), where the CPT and Chat Distillation models again demonstrate superior performance. For instance, the CPT distillation model scores 86% on the battery reasoning benchmark compared to 78% for the LLaMA 70B Distillation model. Although the CPT Distillation model, which relies solely on continuous pretraining, generally achieves superior performance, there are cases where the Chat Distillation model, our OmniScience Reasoning model, outperforms CPT Distillation (as seen in the Battery Q/A task) or achieves similar performance (as seen on the GPQA Diamond benchmark).

These results confirm that DAPT is essential for both general and specialized scientific reasoning, with consistent gaps between DAPT-based and non-DAPT variants.

4.4 Battery Agent for Molecular Screening

In this section, we demonstrate the adaptability of our OmniScience Reasoning model for battery-specific tasks. We develop a battery agent using our OmniScience Reasoning model to rank molecules as potential electrolyte solvents or additives.

Battery Agent Framework. As shown in Fig. 4, our dual-agent framework pairs a generator (OmniScience Reasoning) that proposes molecular grades with a reflector (GPT-o1) that iteratively refines outputs, supported by a RAG pipeline over scientific literature and a short-term memory module.

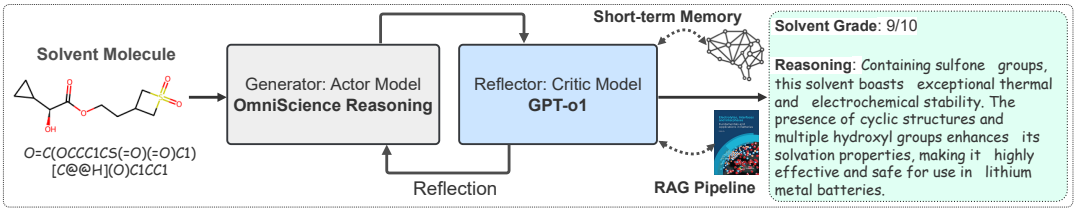

Figure 4: Dual-agent framework for ranking and explaining molecule efficacy as electrolyte solvents or additives. The Generator (OmniScience Reasoning) proposes initial outputs, while the Reflector (GPT-o1) refines and provides feedback autonomously.

Generator	Reflector	Mean Grade	Mean Rank	Hits@5	Hits@10	Hits@20	Hits@50
LLaMA 3.1 70B	GPT-o1	6.4/10	166.70	2/5	3/10	6/20	12/50
OmniScience Chat	GPT-o1	7/10	153.46	4/5	4/10	8/20	15/50
OmniScience Reasoning	**GPT-o1**	**8/10**	**81.26**	**5/5**	**9/10**	**14/20**	**32/50**

Table 4: Battery agent performance for electrolyte solvent ranking. Mean Grade: average grade of 71 known-good solvents (from 700 total). Mean Rank: average position of good molecules in the full ranked list (lower is better). Hits@K: fraction of good molecules in top-K predictions.

Molecular Ranking Results using our Battery Agent Framework. We evaluate the performance of our battery agent (see Table 4) by varying the generator agent while keeping the reflector agent fixed as GPT-o1. The evaluation is conducted on a set of 71 well-known solvent molecules, selected from a diverse pool of approximately 700 molecules that include good, bad, and reasonable electrolyte solvents. We can then sort all 700 molecules by their LLM-assigned grades to assign a unique rank for each molecule. We then compute three key metrics for these 71 molecules to report the quality of our ranking (see Sec. A.3 in the Appendix for further details).

Table 4 shows progressive improvement from LLaMA (Mean Grade 6.4, Mean Rank 166.7) through OmniScience Chat (7.0, 153.5) to OmniScience Reasoning (8.0, 81.3), with Hits@50 nearly tripling from 12/50 to 32/50.

Ablation: Generator-only vs. Dual-agent. To isolate the value of separating *ideation* and *critique*, we compared generator-only systems (o1 or DeepSeek-R1) with and without RAG (short-term memory) to our dual-agent setup (OmniScience generator + o1 reflector). As shown in Table 5, Generator-only agents produce competitive drafts but underperform on ranking fidelity and trace robustness; the dual-agent framework improves mean grade and rank while preserving traceable rationales.

Error analysis. We observe three recurring failure modes: (1) *Table parsing*—the model sometimes misreads multi-column or merged-cell tables, leading to row/column swaps and

Setup	RAG	Mean Grade	Mean Rank	Hits@5	Hits@10	Hits@20
o1 (Gen. only)	Yes	7.6/10	118.4	4/5	6/10	10/20
DeepSeek-R1 (Gen. only)	Yes	7.3/10	126.3	4/5	5/10	8/20
o1 (Gen. only)	No	6.8/10	157.4	4/5	4/10	7/20
DeepSeek-R1 (Gen. only)	No	6.4/10	164.3	3/5	4/10	6/20
OmniScience (Gen.) + o1 (Ref.)	**No**	**7.2 / 10**	**138.6**	**4 / 5**	**5 / 10**	**8 / 20**

Table 5: **Battery molecule ranking: generator-only vs. dual-agent.** Comparison of generator-only baselines (o1, DeepSeek-R1) with/without retrieval short-term memory (RAG) against a dual-agent setup (OmniScience as generator + o1 as reflector). Metrics include *Mean Grade* (1–10; higher is better), *Mean Rank* (lower is better), and *Hits@k* (count of desired molecules in the top k). Under identical prompts/decoding and no external retrieval for the dual-agent, splitting *ideation* and *critique* improves mean grade and ranking fidelity over generator-only without RAG and narrows the gap to RAG-assisted generators.

misassociated numeric values; (2) *Edge electrochemistry cases*—out-of-distribution chemistries (e.g., unusual solvent miscibility or borderline electrochemical-stability windows) trigger either overconfident extrapolations or hedged non-answers; and (3) *Long-context distractors*-with dense literature snippets, the model may cite a nearby but only semantically adjacent sentence.

Mitigations in our production agent include citation-RAG with source highlighting, self-consistency voting, and human-in-the-loop review for high-impact recommendations.

5 Conclusion and Future Work

We introduce OmniScience, a 70B reasoning model that achieves state-of-the-art performance on scientific tasks among similarly-sized models. By combining domain adaptive pretraining with reasoning-based distillation, our approach significantly outperforms the base LLaMA model with only ∼1% additional pretraining compute. Ablations confirm both components are essential—neither DAPT nor reasoning distillation alone matches their combination. OmniScience Reasoning outperforms nearly all non-reasoning models and attains competitive results with much larger reasoning models on GPQA-Diamond, while exceeding all baselines on battery-specific tasks.

Future work will explore domain-specific reasoning distillation and reinforcement learning on reasoning-distilled models (Guo et al., 2025) to further improve specialized task performance.

References

Anthropic. Claude 3.5 sonnet. https://www.anthropic.com/, 2023. Large language model. Accessed: 2025-03-17.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*, 2024.

Sébastien Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Lisa Anne Hendricks, Catalina Dadarlat, Daniel Ziegler, Chris Jones, Albin Cassirer, Andy

Brock, Akbir Khan, Vlad Firoiu, Oriol Vinyals, Andrew Trask, Nicolas Carrazza, Matthew Botvinick, and David Choi. Improving language models by retrieving from trillions of tokens, 2021. arXiv preprint arXiv:2112.04426.

Andrei Z Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pp. 21–29. IEEE, 1997.

Z. Chen, S. Chen, Y. Ning, Q. Zhang, B. Wang, B. Yu, and H. Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*, 2024. URL https://arxiv.org/abs/2410.05080.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022.

José Cassio dos Santos Junior, Rachel Hu, Richard Song, and Yunfei Bai. Domain-driven llm development: Insights into rag and fine-tuning practices. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6416–6417, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Y. Fang, X. Liang, N. Zhang, K. Liu, R. Huang, Z. Chen, and H. Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *arXiv preprint arXiv:2306.08018*, 2023. URL https://arxiv.org/abs/2306.08018.

Yuyuan Feng, Guosheng Hu, and Zhihong Zhang. Gpt4battery: An llm-driven framework for adaptive state of health estimation of raw li-ion batteries. *arXiv preprint arXiv:2402.00068*, 2024.

Nathan Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gomez-Bombarelli, Connor Coley, and Vijay Gadepally. Chemgpt: Neural scaling of deep chemical models. *ChemRxiv*, 2022. doi: 10.26434/chemrxiv-2022-3s512. URL https://chemrxiv.org/engage/chemrxiv/article-details/63c6e3a2d53b6c4e8b7e6e0f.

Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases (VLDB)*, pp. 518–529. VLDB Endowment, 1999.

Frédéric Guillaume and Jacques Rougemont. Nemo: an evolutionary and population genetics programming framework. *Bioinformatics*, 22(20):2556–2557, 2006.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

Gautier Izacard, Lucas Hosseini, Fabio Petroni, Patrick Lewis, Ledell Wu, Wen tau Yih, Sonal Gupta, Nicola De Cao, Antoine Bordes, Stuart Shieber, Sebastian Riedel, and Douwe Kiela. Atlas: Few-shot learning with retrieval augmented language models, 2022. arXiv preprint arXiv:2208.03299.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2020. arXiv preprint arXiv:2005.11401.

Juncai Li and Xiaofei Jiang. Mol-bert: An effective molecular representation with bert for molecular property prediction. *Complexity*, pp. 1–11, 2021. doi: 10.1155/2021/7181815. URL https://doi.org/10.1155/2021/7181815.

Mingjie Liu, Teodor-Dumitru Ene, Robert Kirby, Chris Cheng, Nathaniel Pinckney, Rongjian Liang, Jonah Alben, Himyanshu Anand, Sanmitra Banerjee, Ismet Bayraktaroglu, et al. Chipnemo: Domain-adapted llms for chip design. *arXiv preprint arXiv:2311.00176*, 2023.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://www.aclweb.org/anthology/2020.acl-main.447.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

OpenAI. ChatGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, November 2022. [Accessed: 2023-03-04].

OpenAI. ChatGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, 2024.

OpenAI. ChatGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Junhong Shen, Neil Tenenholtz, James Brian Hall, David Alvarez-Melis, and Nicolo Fusi. Tag-llm: Repurposing general-purpose llms for specialized domains. *arXiv preprint arXiv:2402.05140*, 2024.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Yingheng Tang, Wenbin Xu, Jie Cao, Jianzhu Ma, Weilu Gao, Steve Farrell, Benjamin Erichson, Michael W Mahoney, Andy Nonaka, and Zhi Yao. Matterchat: A multi-modal llm for material science. *arXiv preprint arXiv:2502.13107*, 2025.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2025.

Unstructured.IO. Unstructured: A python library for processing unstructured data. https://github.com/Unstructured-IO/unstructured, 2022. Accessed: February 27, 2025.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models, 2024.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

xAI. Grok 3 beta — the age of reasoning agents. https://x.ai/news/grok-3, 2025.

Aybars Yunusoglu, Dexter Le, Karn Tiwari, Murat Isik, and I Dikmen. Battery state of health estimation using llm framework. *arXiv preprint arXiv:2501.18123*, 2025.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024.

Qiang Zhang, Keyan Ding, Tianwen Lv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. Scientific large language models: A survey on biological & chemical domains. *ACM Computing Surveys*, 57(6):1–38, 2025.

A Appendix

A.1 Training Details

Domain Adaptive Pretraining. Training and validation losses decline rapidly and stabilize without overfitting (Fig. 5), demonstrating effective domain knowledge acquisition.

SFT & Distillation. SFT loss curves (Fig. 6) show efficient adaptation to instruction data. Reasoning distillation on s1K (Muennighoff et al., 2025) achieves near-zero training loss (Fig. 7); following standard practice, the small dataset is used entirely for training with non-independent validation.

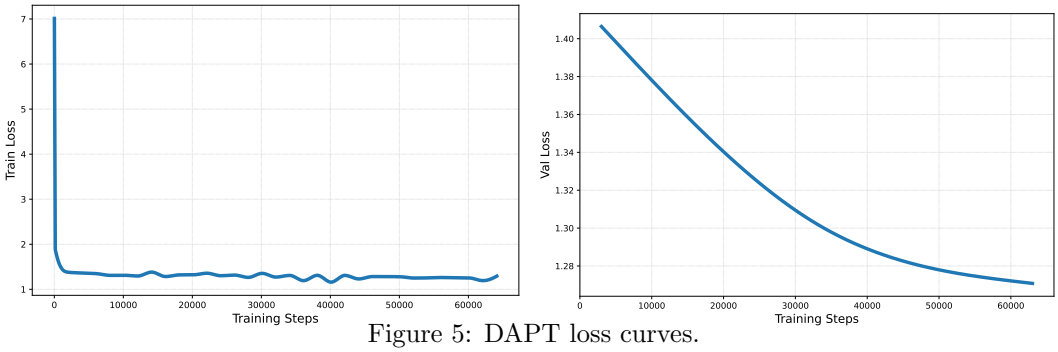

Figure 5: DAPT loss curves.

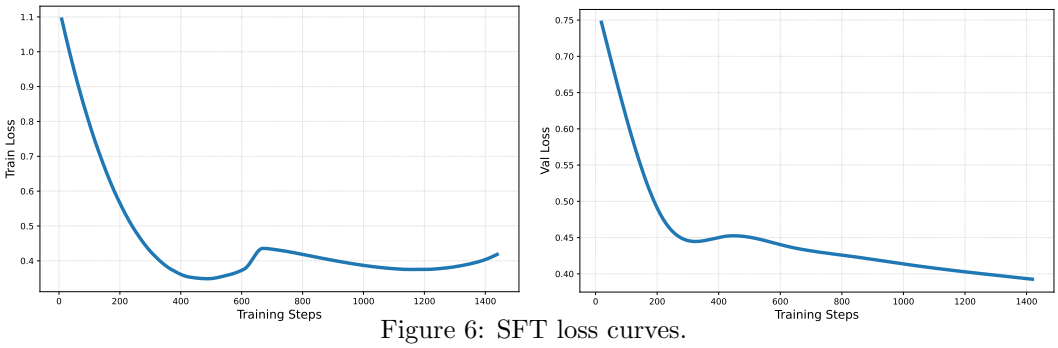

Figure 6: SFT loss curves.

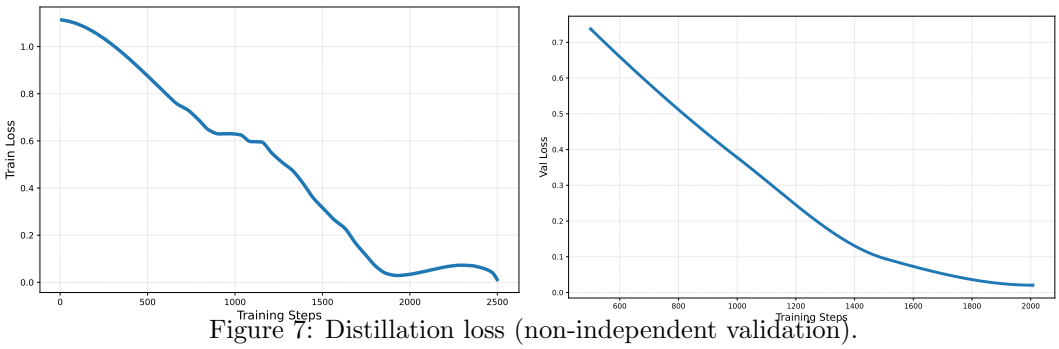

Figure 7: Distillation loss (non-independent validation).

A.2 Pretraining Data

Our 35B-token corpus draws from peer-reviewed journals (3.6M docs), arXiv (1.4M), ChemRxiv (26K), open research (12M), PubChem (60K), academic books (80), and PLOS (200K), ensuring

broad scientific domain coverage.

A.3 Evaluation Metrics

Mean Grade: Average of molecule grades (1-10 scale). **Mean Rank:** Average position of good molecules in ranked list (lower is better). **Hits@k:** Fraction of desired molecules in top-k predictions.

A.4 Additional Benchmarks

We evaluate on *Mol-Instructions* (Fang et al., 2023) (biomolecular text understanding: entity recognition, disease/protein interactions, fact verification) and *ScienceAgentBench* (Chen et al., 2024) (agentic scientific reasoning, 100-item subset), both under unified non-RAG policy.

Model	Entity	Disease	Protein	True/False
Claude 3.7	0.72	0.44	0.19	0.42
Gemini 2.0	0.70	0.43	0.15	0.35
o1	0.65	0.36	0.12	0.44
DeepSeek-R1	0.64	0.41	0.17	0.36
OmniScience-R	**0.66**	**0.38**	**0.15**	**0.36**
Llama 3.1 70B	0.44	0.24	0.06	0.28

Table 6: Mol-Instructions results (unified non-RAG evaluation).

Model	Accuracy
Claude 3.7	0.72
Gemini 2.0	0.72
o1	0.68
DeepSeek-R1	0.66
OmniScience-R	**0.64**
Llama 3.1 70B	0.48

Table 7: ScienceAgentBench (100 items, unified evaluation).

A.5 Example Responses

We present abbreviated examples demonstrating OmniScience's reasoning capabilities. Full examples with complete traces are available in our online repository.

Q1: Calculate expected GED/VED for LIBs with 30% Si-content and pure Si anodes, given 12% Si achieves 320 Wh/kg and 600 Wh/L.[2]

Answer (condensed): For 30% Si: GED ≈ 582 Wh/kg, VED ≈ 294 Wh/L. For pure Si: GED ≈ 1,626 Wh/kg, VED ≈ 3,050 Wh/L. The model systematically derives these via capacity calculations (Si: 4200 mAh/g, graphite: 372 mAh/g), density considerations ($\rho_{\text{Si}} = 2.33$ g/cm^3, $\rho_{\text{graphite}} = 2.09$ g/cm^3), and energy formulas at 3.7 V, acknowledging assumptions yield upper bounds.

[2]Calculations assume constant anode mass/volume for comparability; this overestimates absolute values at higher Si loadings.

Q2: Five 2 V cells (each 0.2 Ω internal resistance) in series, connected to 14 Ω load. Find current: (a) 10 A, (b) 1.4 A, (c) 1.5 A, (d) 2/3 A.

Answer (condensed): Total voltage: $5 \times 2 = 10$ V. Total resistance: $5 \times 0.2 + 14 = 15$ Ω. Current: $I = 10/15 = 2/3$ A. **Answer: (d).** The reasoning trace verifies series voltage addition, internal resistance summation, and Ohm's law application.

A.6 Released Artifacts

Public weights: AWQ 4-bit (g=64) checkpoint at https://huggingface.co/wikiviggy123/OmniScience (subfolder `awq-4bit-g64`, 39 GB). Fits on single 48 GB GPU or two 24 GB GPUs via `device_map="balanced"`. Inference (weights + 8k KV cache) uses ≈42 GB VRAM.

Quick start:

```
from transformers import AutoTokenizer
from awq import AutoAWQForCausalLM

repo = "wikiviggy123/OmniScience"
subfolder = "awq-4bit-g64"
tok = AutoTokenizer.from_pretrained(repo, subfolder=subfolder)
model = AutoAWQForCausalLM.from_quantized(
    repo, subfolder=subfolder, device_map="auto", fuse_layers=True)

prompt = "Explain the Nernst equation in one paragraph."
out = model.generate(**tok(prompt, return_tensors="pt").to(model.device),
                     max_new_tokens=4096)
print(tok.decode(out[0], skip_special_tokens=True))
```

Repository includes config, tokenizer, and safetensors shards for reproducibility.