Olmo, Molmo, and Open Multimodal AI Infrastructure to Accelerate Science

Noah A. Smith

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Ai2

UWNLP

AI isn’t just chatbots, agents, and APIs

Ai2

Science-serving AI needs *infrastructure* that scientists can inspect and control.

Model flows are how we make that possible.

Model Flow Includes …

For every training stage:

- All training data
- Model weights
- Intermediate checkpoints
- Detailed recipes
- Code to reproduce
- Open evaluations
- Documentation and analysis

Full openness; everything needed for reproducibility!

Olmo 3

Olmo Team*

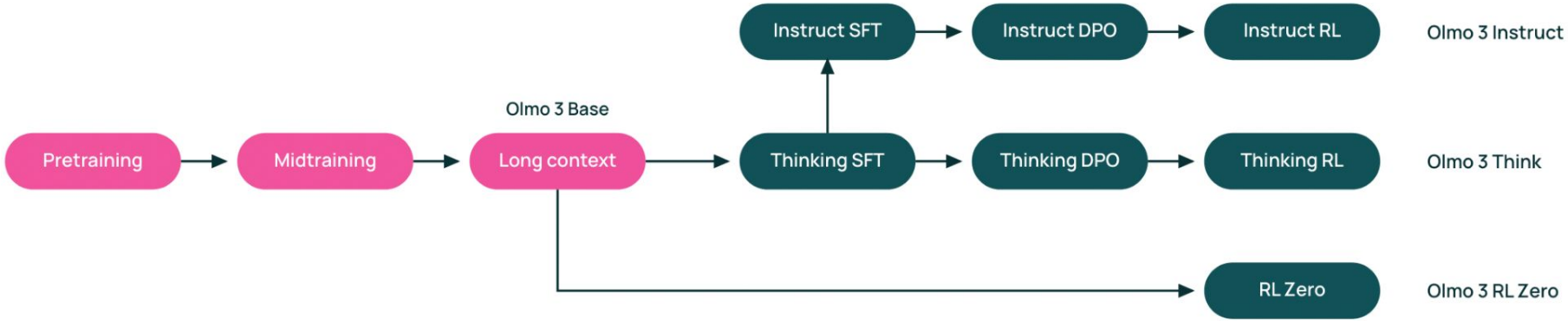
Allyson Ettinger[♥1] Amanda Bertsch[♥1,3] Bailey Kuehl[♥1] David Graham[♥1]
David Heineman[♥1] Dirk Groeneveld[♥1] Faeze Brahman[♥1] Finbarr Timbers[♥1]
Hamish Ivison[♥1,2] Jacob Morrison[♥1,2] Jake Poznanski[♥1] Kyle Lo[♥1,2] Luca Soldaini[♥1]
Matt Jordan[♥1] Mayee Chen[♥1,4] Michael Noukhovitch[♥1,5,6] Nathan Lambert[♥1]
Pete Walsh[♥1] Pradeep Dasigi[♥1] Robert Berry[♥1] Saumya Malik[♥1] Saurabh Shah[♥1]
Scott Geng[♥1,2] Shane Arora[♥1] Shashank Gupta[♥1] Taira Anderson[♥1] Teng Xiao[♥1]
Tyler Murray[♥1] Tyler Romero[♥1] Victoria Graf[♥1,2]

Akari Asai[1,3] Akshita Bhagia[1] Alexander Wettig[7] Alisa Liu[2] Aman Rangapur[1]
Chloe Anastasiades[1] Costa Huang[1] Dustin Schwenk[1] Harsh Trivedi[1] Ian Magnusson[1,2]
Jaron Lochner[1] Jiacheng Liu[1] Lester James V. Miranda[1] Maarten Sap[1,3] Malia Morgan[1]
Michael Schmitz[1] Michal Guerquin[1] Michael Wilson[1] Regan Huff[1] Ronan Le Bras[1]
Rui Xin[2] Rulin Shao[2] Sam Skjonsberg[1] Shannon Zejiang Shen[8] Shuyue Stella Li[2]
Tucker Wilde[1] Valentina Pyatkin[1] Will Merrill[1] Yapei Chang[9] Yuling Gu[1] Zhiyuan Zeng[1,2]

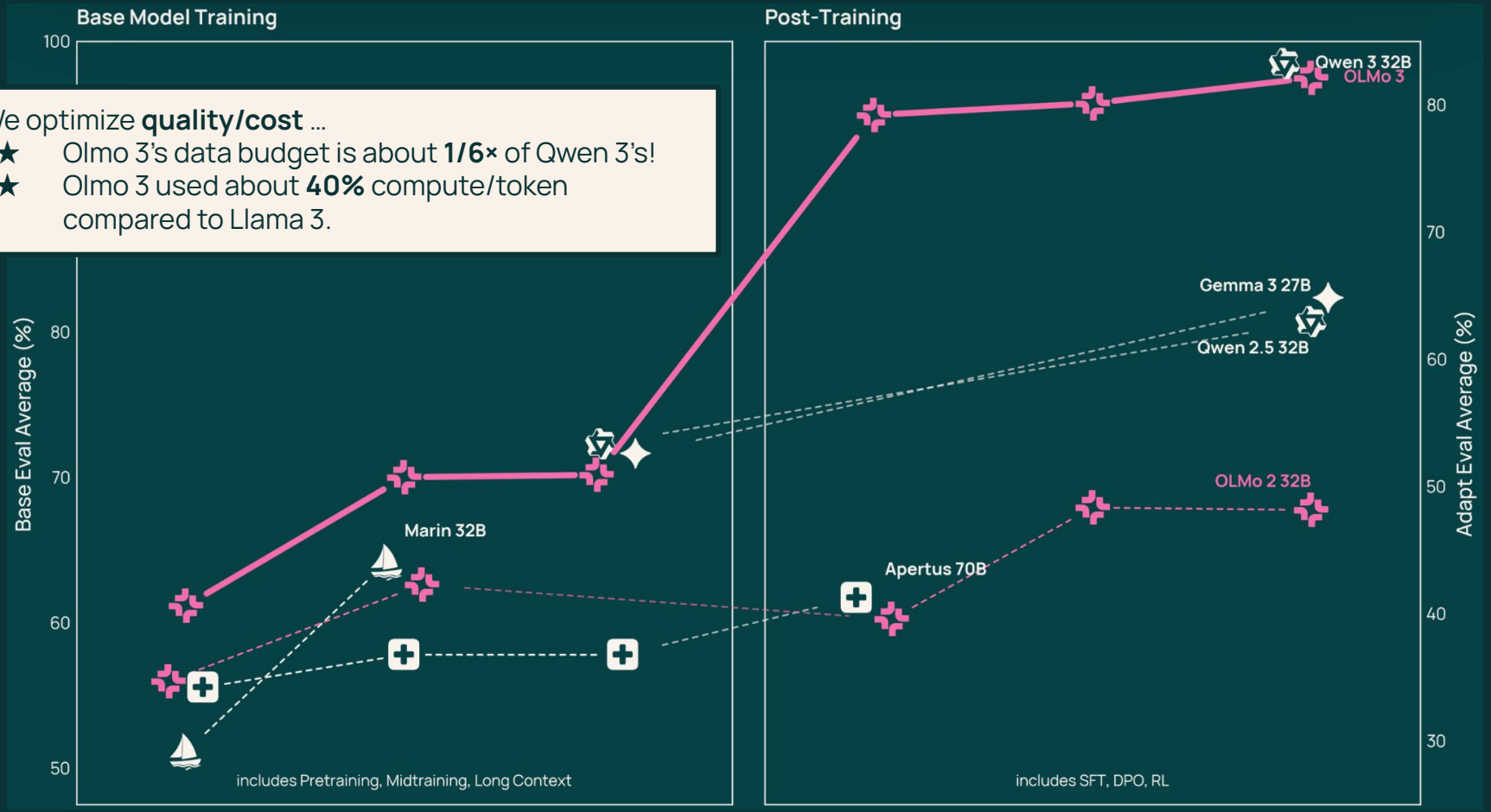
Ashish Sabharwal[1] Luke Zettlemoyer[2] Pang Wei Koh[1,2]
Ali Farhadi[1,2] Noah A. Smith[♥1,2] Hannaneh Hajishirzi[♥1,2]

[1]Allen Institute for AI [2]University of Washington [3]Carnegie Mellon University [4]Stanford University [5]Mila
[6]Université de Montréal [7]Princeton University [8]Massachusetts Institute of Technology [9]University of Maryland

Example: Olmo 3 = the whole model flow!



Demo:
https://playground.allenai.org/


Base Model Training
Post-Training
We optimize **quality/cost** …
★ Olmo 3's data budget is about **1/6×** of Qwen 3's!
★ Olmo 3 used about **40%** compute/token compared to Llama 3.
Base Eval Average (%)
Adapt Eval Average (%)
100
80
70
60
50
40
30
Qwen 3 32B
OLMo 3
Gemma 3 27B
Qwen 2.5 32B
OLMo 2 32B
Marin 32B
Apertus 70B
includes Pretraining, Midtraining, Long Context
includes SFT, DPO, RL

August 2025: $152M in support from NSF and NVIDIA!

Open Multimodal AI Infrastructure to Accelerate Science

What do scientists want?

- Literature review & information retrieval
- Data processing, analysis, & code generation
- Research workflow & bureaucracy
- Quality, reasoning, & nuance
- Specialized and domain-specific tools
- Document & image data parsing
- Multimodal & heterogeneous datasets
- Integration from multiple data sources
- Large, high-dimensional, & tabular data
- Specialized data & contextual interpretation

Take the survey!

Ai2

Ground in data

- Literature review & information retrieval
- Data processing, analysis, & code generation
- Research workflow & bureaucracy
- Quality, reasoning, & nuance
- Specialized and domain-specific tools

- Document & image data parsing
- Multimodal & heterogeneous datasets
- Integration from multiple data sources
- Large, high-dimensional, & tabular data
- Specialized data & contextual interpretation

Take the survey!

Communicate through literature

- Literature review & information retrieval
- Data processing, analysis, & code generation
- Research workflow & bureaucracy
- Quality, reasoning, & nuance
- Specialized and domain-specific tools

- Document & image data parsing
- Multimodal & heterogeneous datasets
- Integration from multiple data sources
- Large, high-dimensional, & tabular data
- Specialized data & contextual interpretation

Take the survey!

Ai2

We are approaching these challenges through R&D across the whole model flow.

Current Portfolio Tour

1. Molmo 2 and beyond: control over perception and grounding
2. DR Tülu: control over long-form research workflows
3. Olmo Hybrid: improving efficiency tradeoffs
4. Hidden cost of thinking
5. Branch-Adapt-Route: control over specialization

Current Portfolio Tour

1. **Molmo 2 and beyond: control over perception and grounding**
2. DR Tülu: control over long-form research workflows
3. Olmo Hybrid: improving efficiency tradeoffs
4. Hidden cost of thinking
5. Branch-Adapt-Route: control over specialization

Molmo and PixMo (CVPR 2025 best paper honorable mention):

open multimodal competence depends on open data, including detailed human-generated captions, instruction data, and pointing to localize visual evidence, bridging to agents/robots/web systems that need to act

Demo:
https://playground.allenai.org/

Molmo's impact on industry-wide image pointing performance

Ai2

Molmo2

Open Weights and Data for Vision-Language Models with Video Understanding and Grounding

(CVPR 2026)

Christopher Clark[♥1*] Jieyu Zhang[♥1,2*] Zixian Ma[♥1,2*] Jae Sung Park[♥1,2*] Mohammadreza Salehi[♥1,2] Rohun Tripathi[♥1] Sangho Lee[♥1]

Zhongzheng Ren[1,2] Chris Dongjoo Kim[1] Yinuo Yang[2] Vincent Shao[2] Yue Yang[1] Weikai Huang[2] Ziqi Gao[1] Taira Anderson[1] Jianrui Zhang[1] Jitesh Jain[1] George Stoica[1] Winson Han[1]

Ali Farhadi[1,2] Ranjay Krishna[♥1,2]

[1]Allen Institute for AI, [2]University of Washington

Demo:
https://playground.allenai.org/

Molmo2

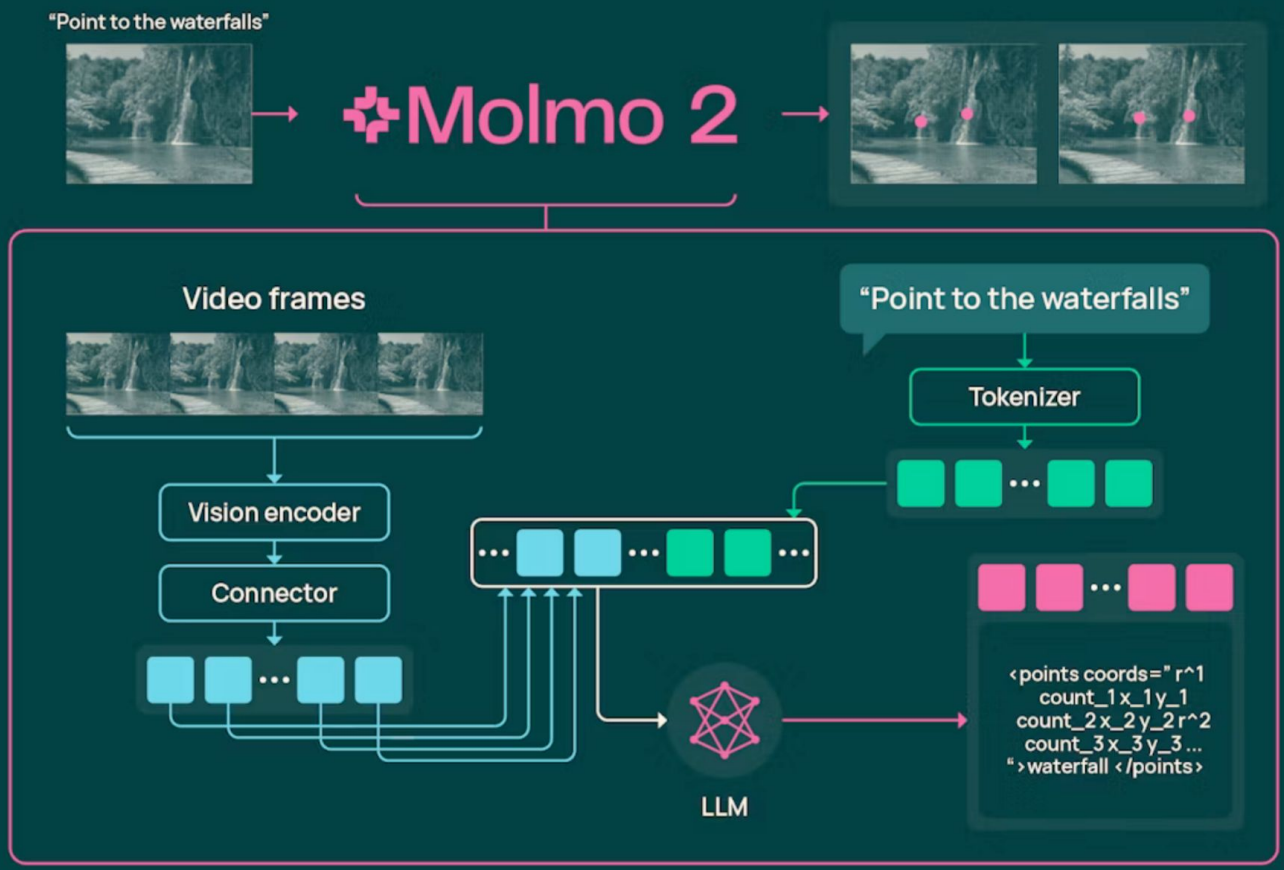
Open Weights and Data for Vision-Language Models with Video Understanding and Grounding

(CVPR 2026)

- **Molmo2:** supports single images, multi-image inputs, and video, and can produce free-form language plus grounded outputs such as spatio-temporal points, object tracks, and grounded chains-of-thought
    - 9 new datasets, including video pointing/tracking, dense video captioning, long-form QA, long-video QA, and multi-image datasets
    - open weights/data/code and no distillation from proprietary VLMs

Demo:
https://playground.allenai.org/


Molmo 2


"Point to the waterfalls"
Molmo 2
Video frames
"Point to the waterfalls"
Tokenizer
Vision encoder
Connector
LLM
<points coords=" r^1
count_1 x_1 y_1
count_2 x_2 y_2 r^2
count_3 x_3 y_3 ...
"›waterfall ‹/points›

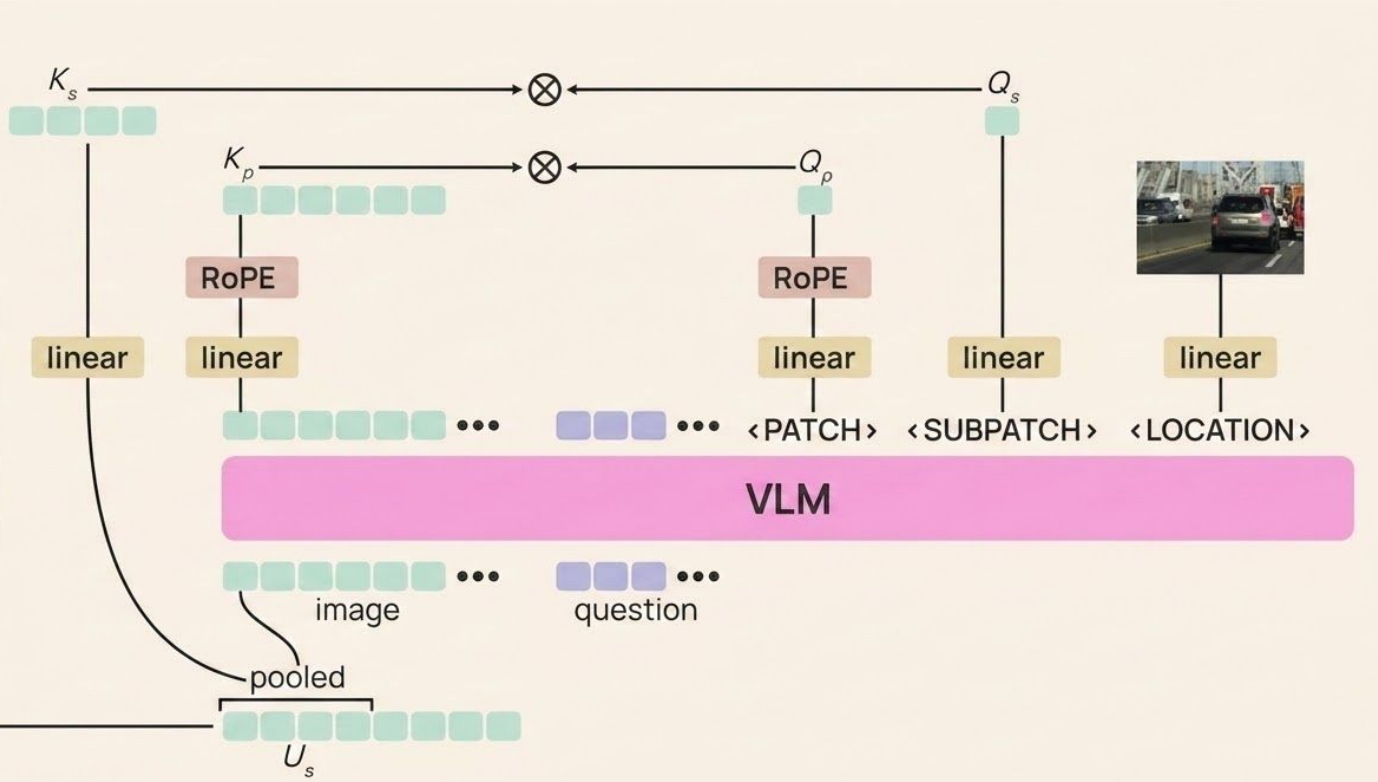
Ai2

MolmoPoint
Better Pointing for VLMs with Grounding Tokens

Christopher Clark[♥1] Yue Yang[♥1] Jae Sung Park[♥1] Zixian Ma[1,2] Jieyu Zhang[1,2] Rohun Tripathi[1]
Mohammadreza Salehi[1,2] Sangho Lee[1] Taira Anderson[1] Winson Han[1] Ranjay Krishna[♥1,2]

[1]Allen Institute for AI, [2]University of Washington

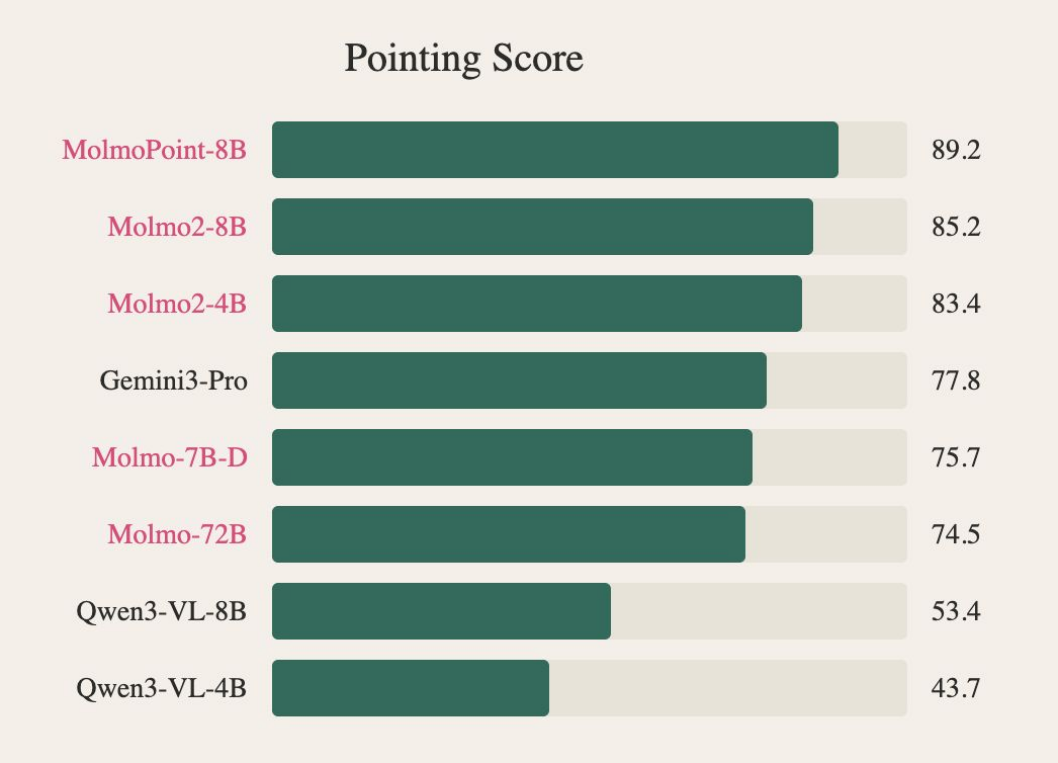
MolmoPoint replaces text-coordinate generation with **grounding tokens** that directly select visual tokens, then refine to subpatch and location. This is a cleaner control surface than asking a language model to emit coordinates as text.

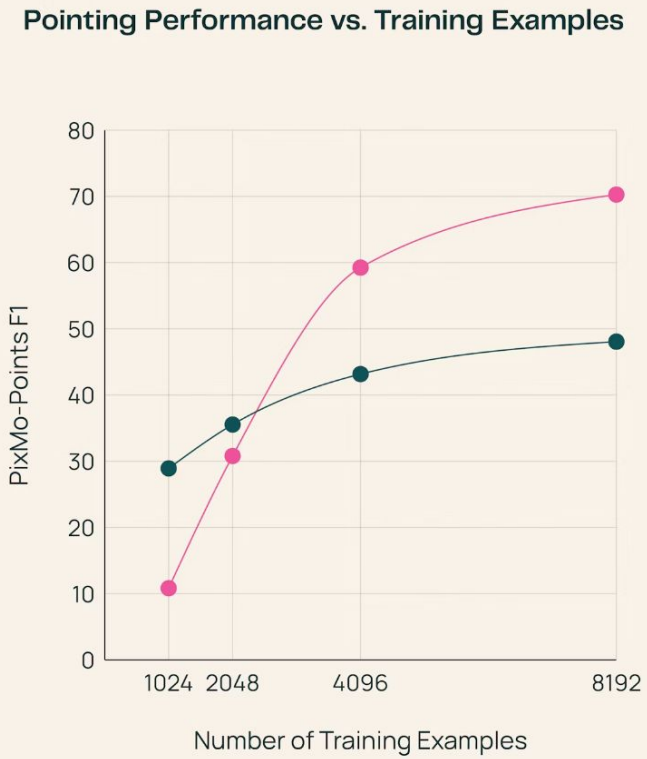



87
CLEARANCE 13-6
Pkwy
3/4 MILE


MolmoPoint
"point to the right brake light of the left most car"

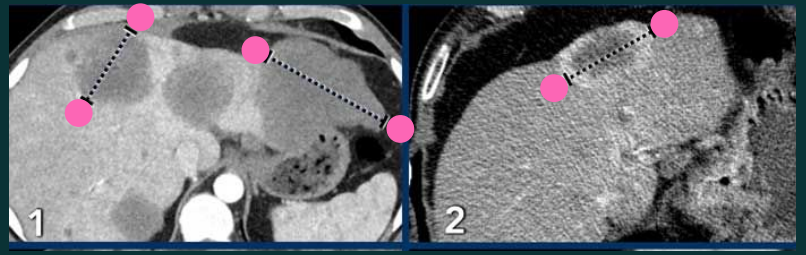
MolmoPoint: Better and More Efficient Grounding





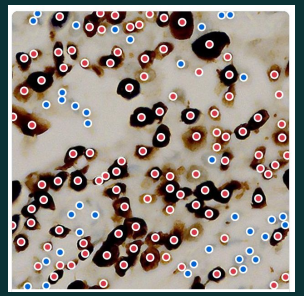
Ai2

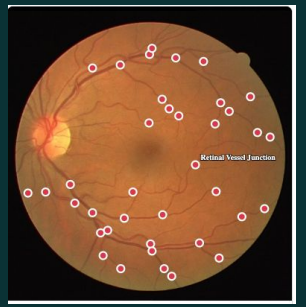
Grounding turns multimodal AI from “trust me” into “look here,” which is essential for science, where answers have to be checked against observable evidence.



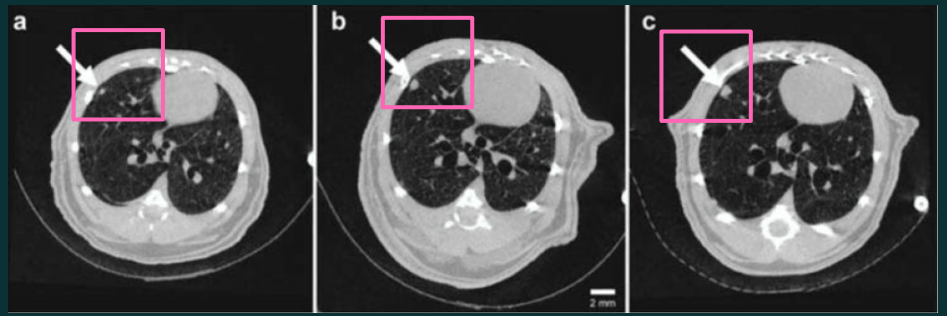
measurement

counting





longitudinal comparison



Ai2

Current Portfolio Tour

1. Molmo 2 and beyond: control over perception and grounding
2. **DR Tülu: control over long-form research workflows**
3. Olmo Hybrid: improving efficiency tradeoffs
4. Hidden cost of thinking
5. Branch-Adapt-Route: control over specialization

DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research (ICML 2026)

Rulin Shao[1♡†], **Akari Asai**[2,3♡†], **Shannon Zejiang Shen**[4♡†], **Hamish Ivison**[1,2♡†], **Varsha Kishore**[1,2†], **Jingming Zhuo**[1†], **Xinran Zhao**[3], **Molly Park**[1], **Samuel Finlayson**[1,5], **David Sontag**[4], **Tyler Murray**[2], **Sewon Min**[2,6], **Pradeep Dasigi**[2], **Luca Soldaini**[2], **Faeze Brahman**[2], **Wen-tau Yih**[1], **Tongshuang Wu**[3], **Luke Zettlemoyer**[1], **Yoon Kim**[4], **Hannaneh Hajishirzi**[1,2], **Pang Wei Koh**[1,2]

[1]University of Washington, [2]Allen Institute for AI, [3]Carnegie Mellon University
[4]Massachusetts Institute of Technology, [5]Seattle Children's Hospital, [6]University of California, Berkeley

Demo: https://www.dr-tulu.org/

DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research (ICML 2026)

- End-to-end recipe for long-form “deep research” agents
- Trainable and inspectable research workflows: agent stack includes search/browsing tools, asynchronous tool use, MCP-based infrastructure, and an evaluation suite
- Synthesis **grounded in external evidence**:
  planning the investigation, searching for relevant sources, integrating evidence across documents, writing answers with attribution, and evaluating citation precision and recall.
- RL with **evolving rubrics**:
  DR needs long-form synthesis where rubrics have to evolve with what the model discovers, not just verifiable rewards

Demo: https://www.dr-tulu.org/

DR Tulu

<call_tool name="web browser">

Planning

Tool Call

Thinking

Tool Call

Thinking

Tool Call

Reflection

Final Answer

Web Browser

Website: Evaluating the latest state of the art models in road surface monitoring technology (2025)

↳ [evaluating website content]

</tool call>

User

How did Netflix manage to successfully adapt One Hundred Years of Solitude, a notoriously difficult book to bring to the screen?

DR Tulu

Agentic Workflow

Think | Tool 1 | Think | Tool 2 | ... | Answer

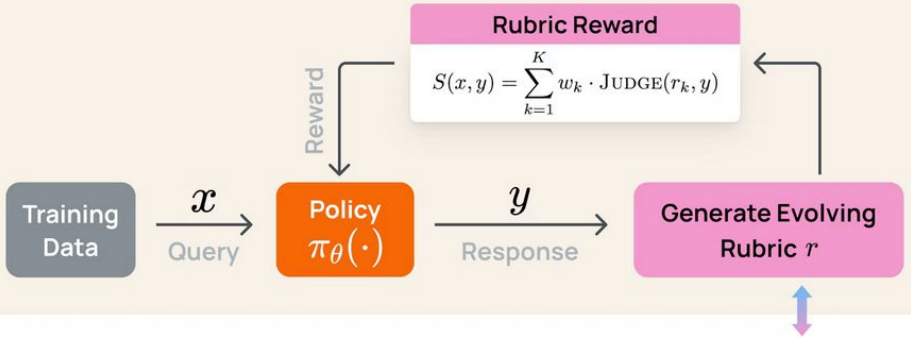
Long-form Report with Citations

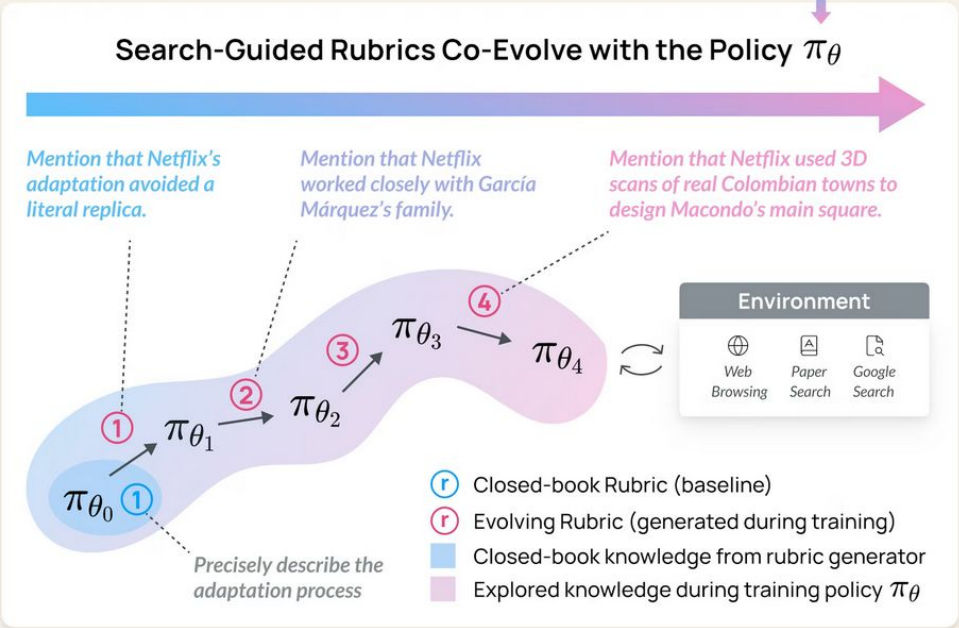
Netflix's adaptation avoided a literal replica of Macondo and instead fused real locations with meticulously built sets [1] to honor the novel's essence while giving the show contemporary visual grammar. The production grounded magical realism in front-of-camera practical craft, relying on makeup, special effects,[2] ..... The location strategy and production design...

Sources

[1] The production team behind Netflix's adaptation of "One Hundred Years of Solitude," LA Times

[2] The article discusses Netflix's adaptation of Gabriel García Márquez's celebrated novel, NY Times





Average Performance Across Deep Research Benchmarks

Current Portfolio Tour

1. Molmo 2 and beyond: control over perception and grounding
2. DR Tülu: control over long-form research workflows
3. **Olmo Hybrid: improving efficiency tradeoffs**
4. Hidden cost of thinking
5. Branch-Adapt-Route: control over specialization

Olmo Hybrid: From Theory to Practice

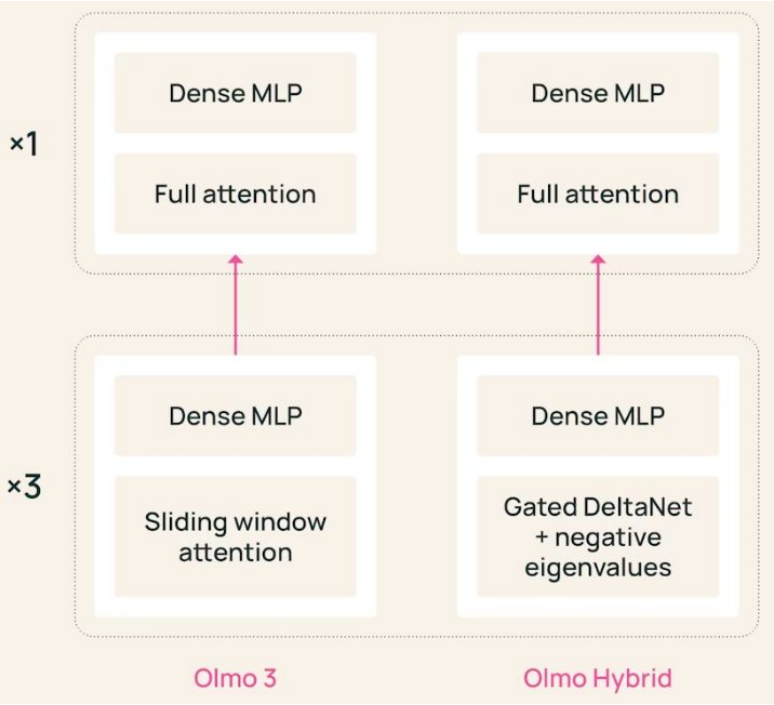
William Merrill[♥1] Yanhong Li[♥1] Tyler Romero[♥1] Anej Svete[♥1,4] Caia Costello[♥3]
Pradeep Dasigi[1] Dirk Groeneveld[1] David Heineman[1] Bailey Kuehl[1] Nathan Lambert[1]
Chuan Li[3] Kyle Lo[1,2] Saumya Malik[1] DJ Matusz[3] Benjamin Minixhofer[1,5]
Jacob Morrison[1,2] Luca Soldaini[1] Finbarr Timbers[1] Pete Walsh[1] Noah A. Smith[1,2]
Hannaneh Hajishirzi[1,2] Ashish Sabharwal[♥1]

(COLM 2026)

[1]Allen Institute for AI [2]University of Washington [3]Lambda ML Team
[4]ETH Zürich [5]University of Cambridge

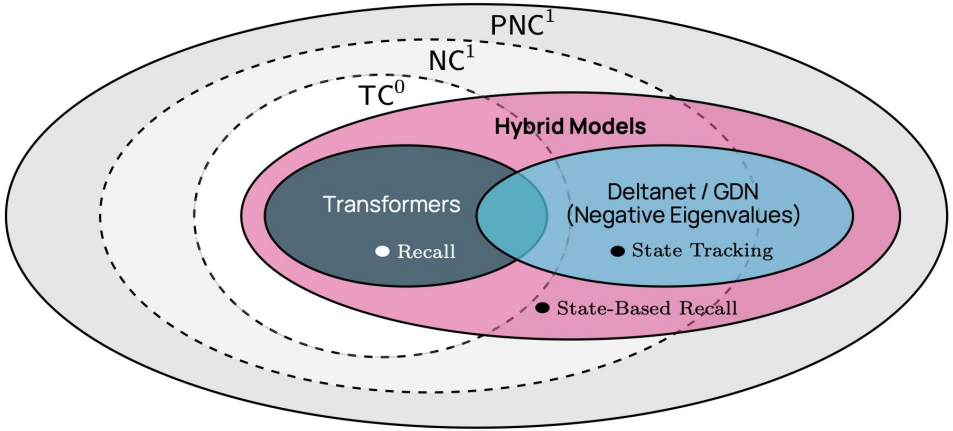
Olmo Hybrid: From Theory to Practice

- Gated DeltaNet (GDN) + attention
- 3:1 hybridization ratio
- 7B parameters, 6T tokens
- Copies Olmo 3 7B in other choices (matched in params & train. throughput)

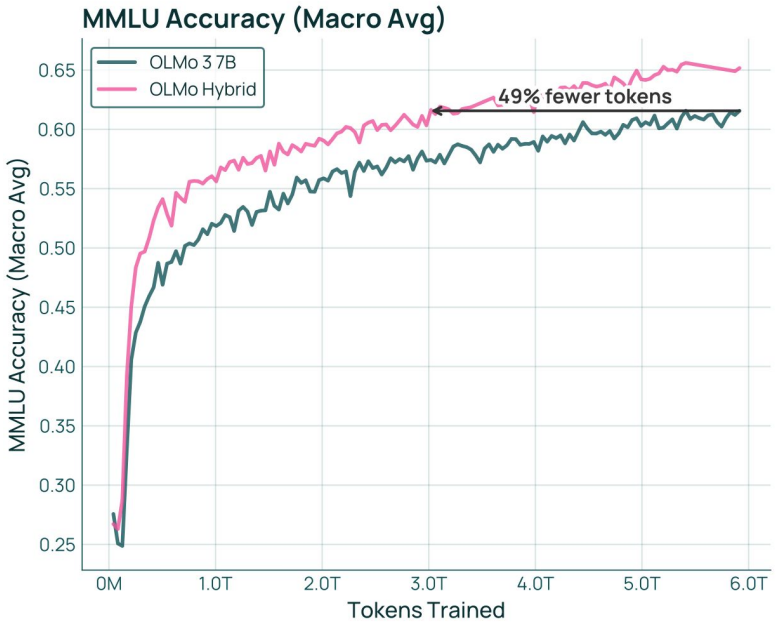


Hybrid models are more expressive than transformers!

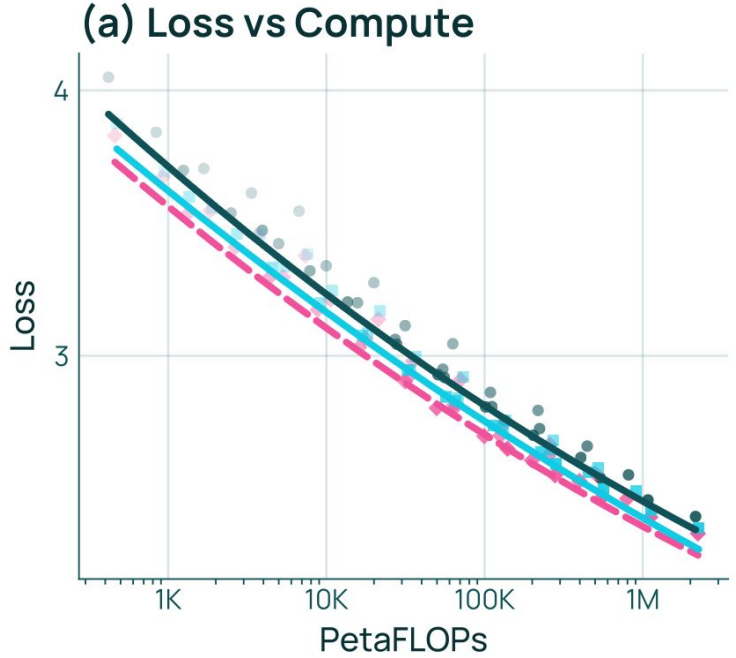
- Transformers excel at **recall** but cannot represent state tracking
- Linear RNNs can represent **state tracking** but struggle at recall
- Hybrid models inherit state tracking and recall
- Moreover, we prove they are more expressive than both transformers and RNNs!



Big win: hybrid models are more token-efficient

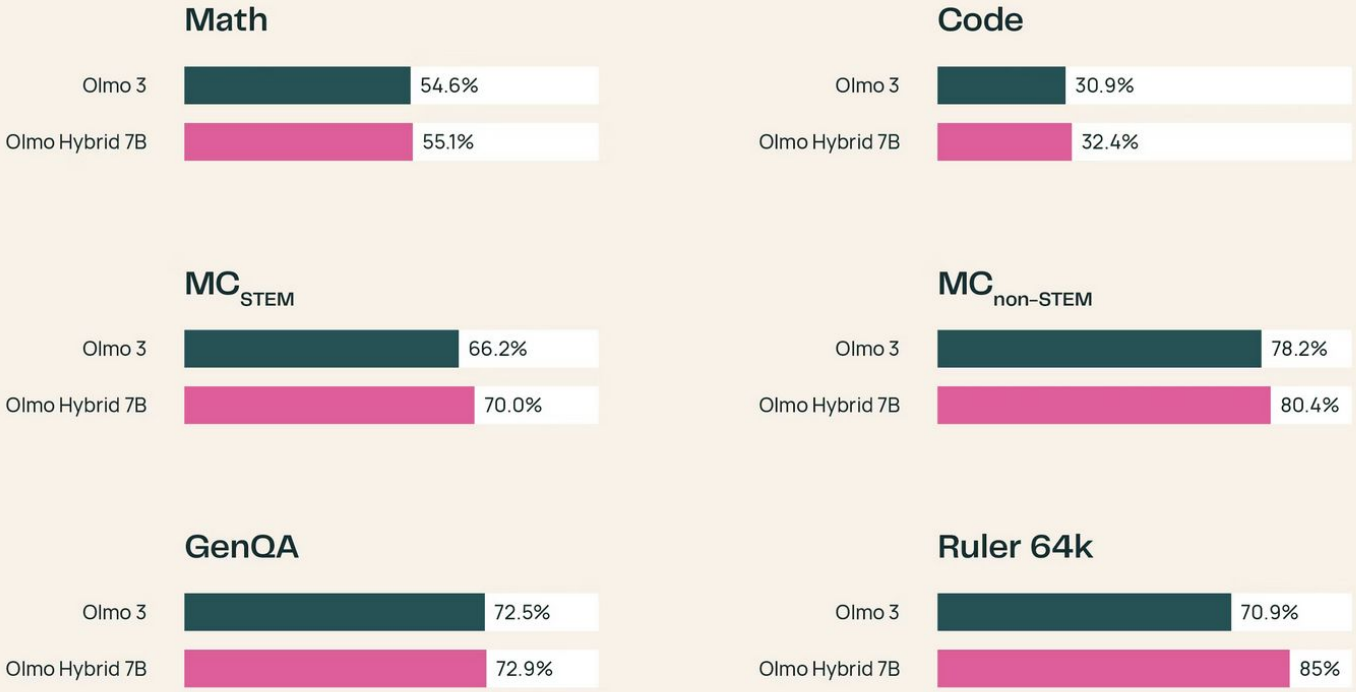


Olmo Hybrid matches Olmo 3 on MMLU with half as many tokens or FLOPs



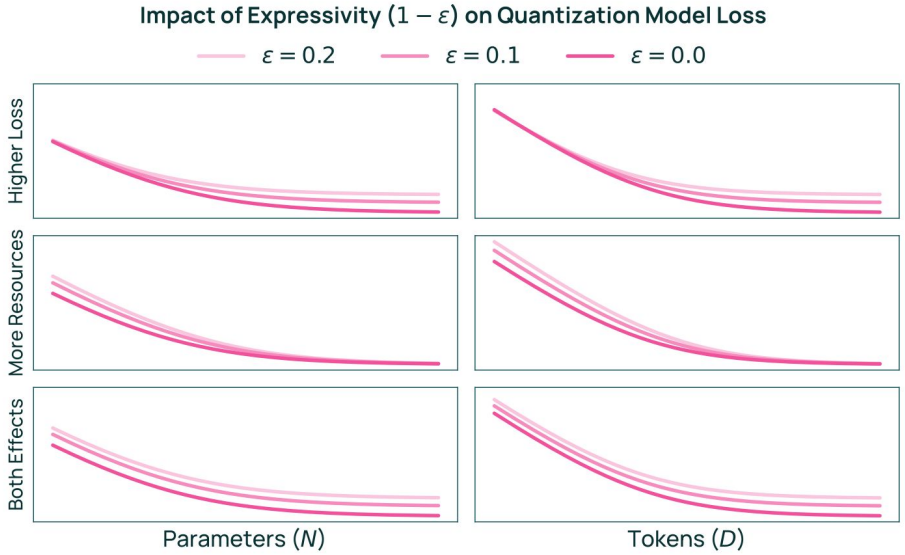
Controlled scaling studies confirm hybrid model is more token/compute-efficient

Olmo Hybrid vs. Olmo 3 (7B)



Expressive power → better scaling

- Open question: does hybrid models' expressive power explain their better scaling in practice?
- Under the quantization model of scaling laws (Michaud et al., 2023), we prove **more expressive models have better scaling laws**
- Thus, the greater expressive power of hybrid models plausibly explains their better scaling



Current Portfolio Tour

1. Molmo 2 and beyond: control over perception and grounding
2. DR Tülu: control over long-form research workflows
3. Olmo Hybrid: improving efficiency tradeoffs
4. **Hidden cost of thinking**
5. Branch-Adapt-Route: control over specialization

The Hidden Cost of Thinking: Energy Use and Environmental Impact of LMs Beyond Pretraining

Jacob Morrison
University of Washington
Allen Institute for AI

Noah A. Smith
University of Washington
Allen Institute for AI

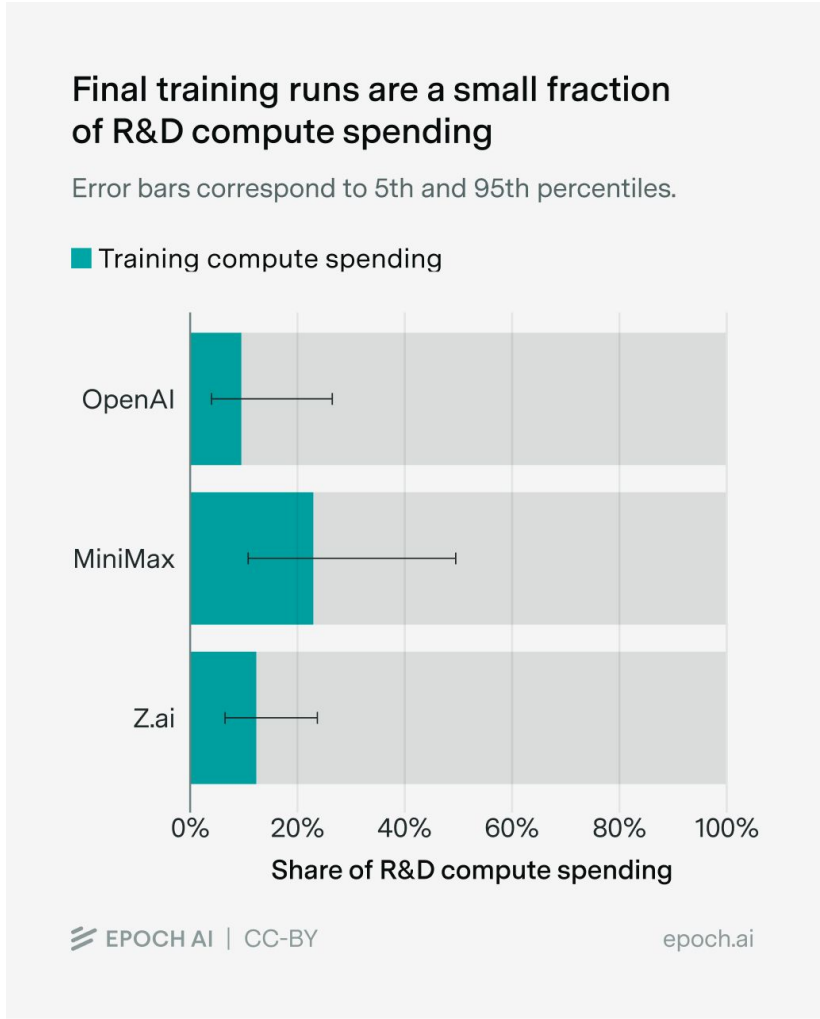
Emma Strubell
Carnegie Mellon University

(COLM 2026)

The Hidden Cost of Thinking: Energy Use and Environmental Impact of LMs Beyond Pretraining

- We cannot understand the cost of AI without looking at the whole model flow, including failed runs, ablations, and experimentation.
- Key findings:
  - **Reasoning is not free.** For OLMo 3 32B, the Think variant's post-training used **17× more datacenter energy** than the Instruct variant's post-training. The driver is RLVR rollout generation.
  - **Water depends on where and how the compute is powered and cooled, not just the workload.** For Olmo 3 training, onsite cooling water was effectively zero because of closed-loop cooling; estimated water use came from power generation.
  - **Development dominates final runs.** OLMo 3 development used **82.2%** of total GPU hours excluding offline data generation. Reporting only final runs understates impact by roughly **5×**.

“Reporting only final runs understates impact by roughly 5×.” – corroborated by Epoch.ai’s analysis (March 2026)



Reasoning, agents, and tool use are control knobs with real costs. *Scientists* should decide when the cost is worth it.

Current Portfolio Tour

1. Molmo 2 and beyond: control over perception and grounding
2. DR Tülu: control over long-form research workflows
3. Olmo Hybrid: improving efficiency tradeoffs
4. Hidden cost of thinking
5. **Branch-Adapt-Route: control over specialization**

Train Separately, Merge Together: Modular Post-Training with Mixture-of-Experts

Jacob Morrison[*1,3] **Sanjay Adhikesaven**[*2,3] **Akshita Bhagia**[3]
Matei Zaharia[2] **Noah A. Smith**[1,3] **Sewon Min**[2,3]

[1]University of Washington [2]University of California, Berkeley [3]Allen Institute for AI

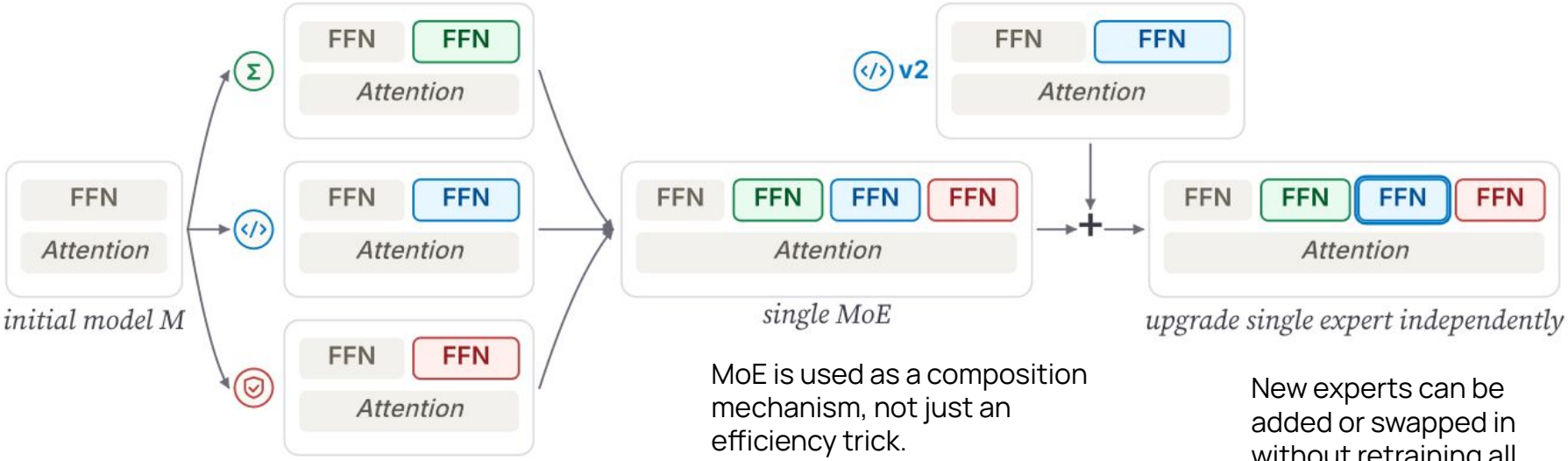
jacobm00@cs.washington.edu sanjay.adhikesaven@berkeley.edu

(COLM 2026)

Train Separately, Merge Together: Modular Post-Training with Mixture-of-Experts

- Updating a model after it's trained is extremely expensive!
- Recipe: **branch, adapt, route (BAR**; train separate domain experts, adapt each through the stages it needs, then compose them with a Mixture-of-Experts architecture and lightweight router training.
- BAR keeps an anchor expert initialized from the original post-trained model, so specialization does not require sacrificing the base model's behavior.
- **BAR makes upgrades cheap and local.**

Branch-Adapt-Route



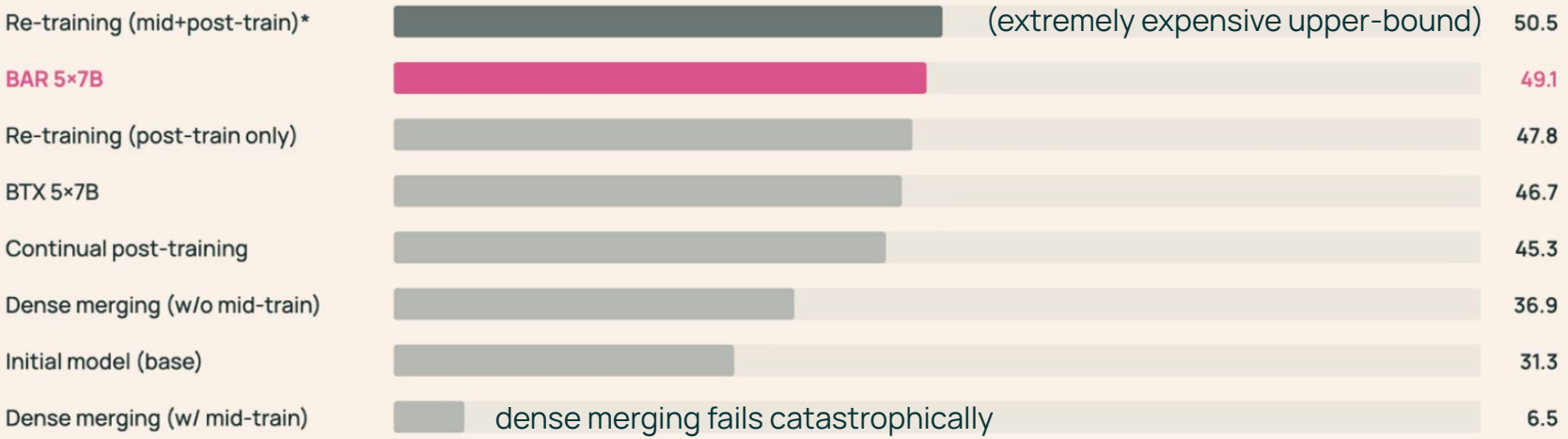
Independent expert training: each domain/capability set can have its own pipeline (e.g., math and code use mid-training → SFT → RLVR, while tool use and safety use SFT only)

MoE is used as a composition mechanism, not just an efficiency trick.

New experts can be added or swapped in without retraining all previous experts; linear rather than quadratic scaling for updates!

Branch-Adapt-Route: Benchmark Results

19 benchmarks (math, code, tool use, safety); all models 7B scale on Olmo 2



Reproducing commercial AI is too small a goal.

Instead, infrastructure should let scientific communities do things the market will not prioritize:

> inspect the system, adapt it to local scientific requirements, study its development, control its costs, and specialize it for long-tail domains.

Let’s build open infrastructure together!

Acknowledgments

- The entire Ai2 team
- Noah's ARK (my students and postdocs at UW)
- Microsoft Azure
- National AI Research Resource
- Oak Ridge Leadership Computing Facility (allocation on Frontier)
- Lambda
- Nvidia
- National Science Foundation

Thank you!