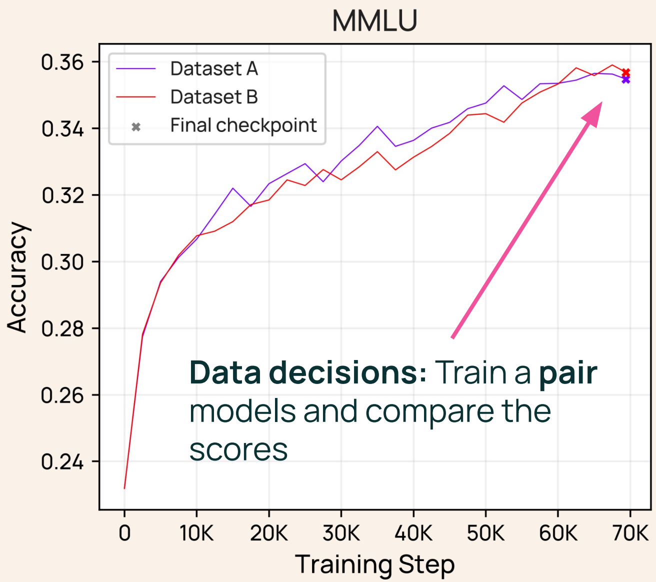
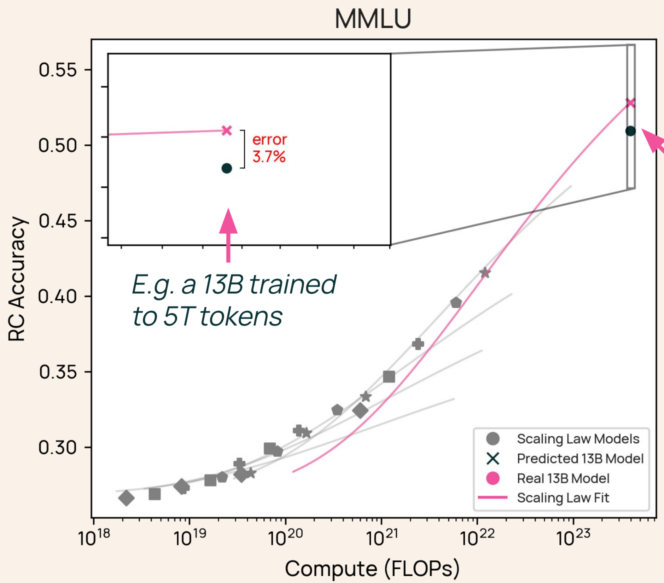
Spotlight poster on Wednesday, Dec 3 at 4:30 in Hall C,D,E

Signal and Noise: A Framework for Reducing Uncertainty in Language Model Evaluation

David Heineman, Valentin Hofmann, Ian Magnusson, Yuling Gu, Noah A. Smith, Hannaneh Hajishirzi, Kyle Lo, Jesse Dodge

Building Language Models means making decisions





Scaling laws: Train many **small** models and extrapolate the performance

DataDecide: How to Predict Best Pretraining Data with Small Experiments (ICML, 2025)
Establishing Task Scaling Laws via Compute-Efficient Model Ladders (COLM, 2025)

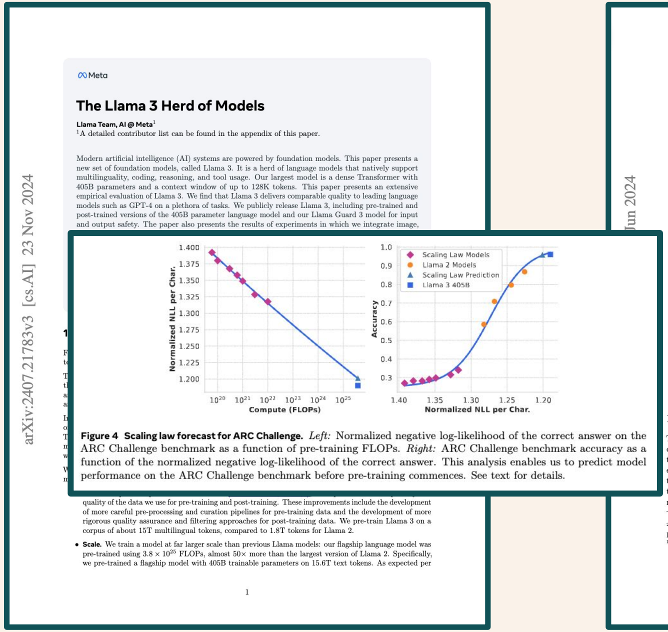
Ai2

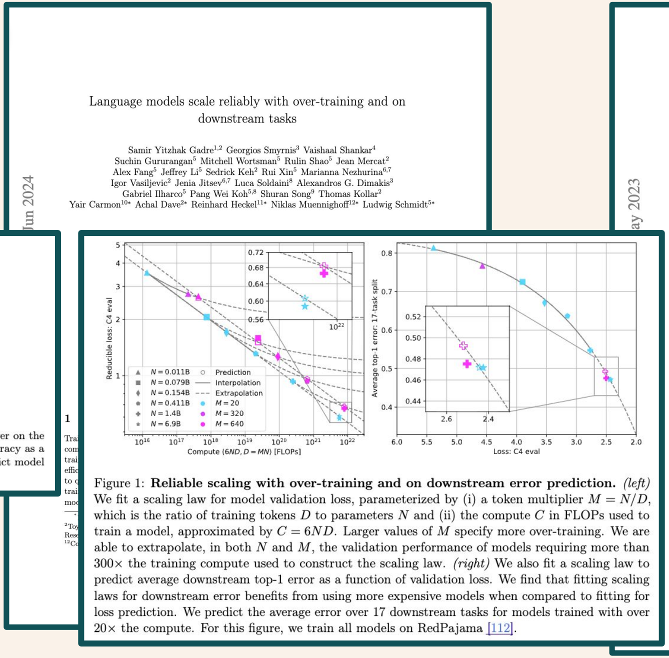
Downstream tasks are now core to building models …

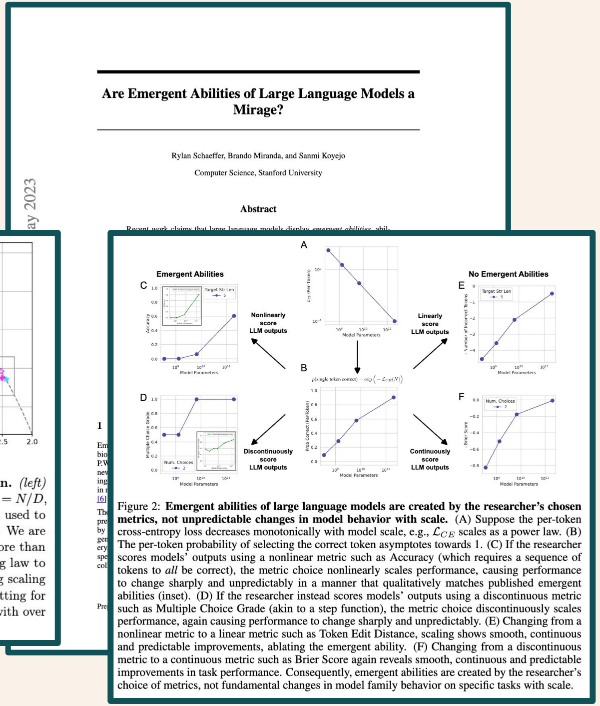
… some tasks are useful!

… some tasks aren't predictable

… some metrics hide real capability



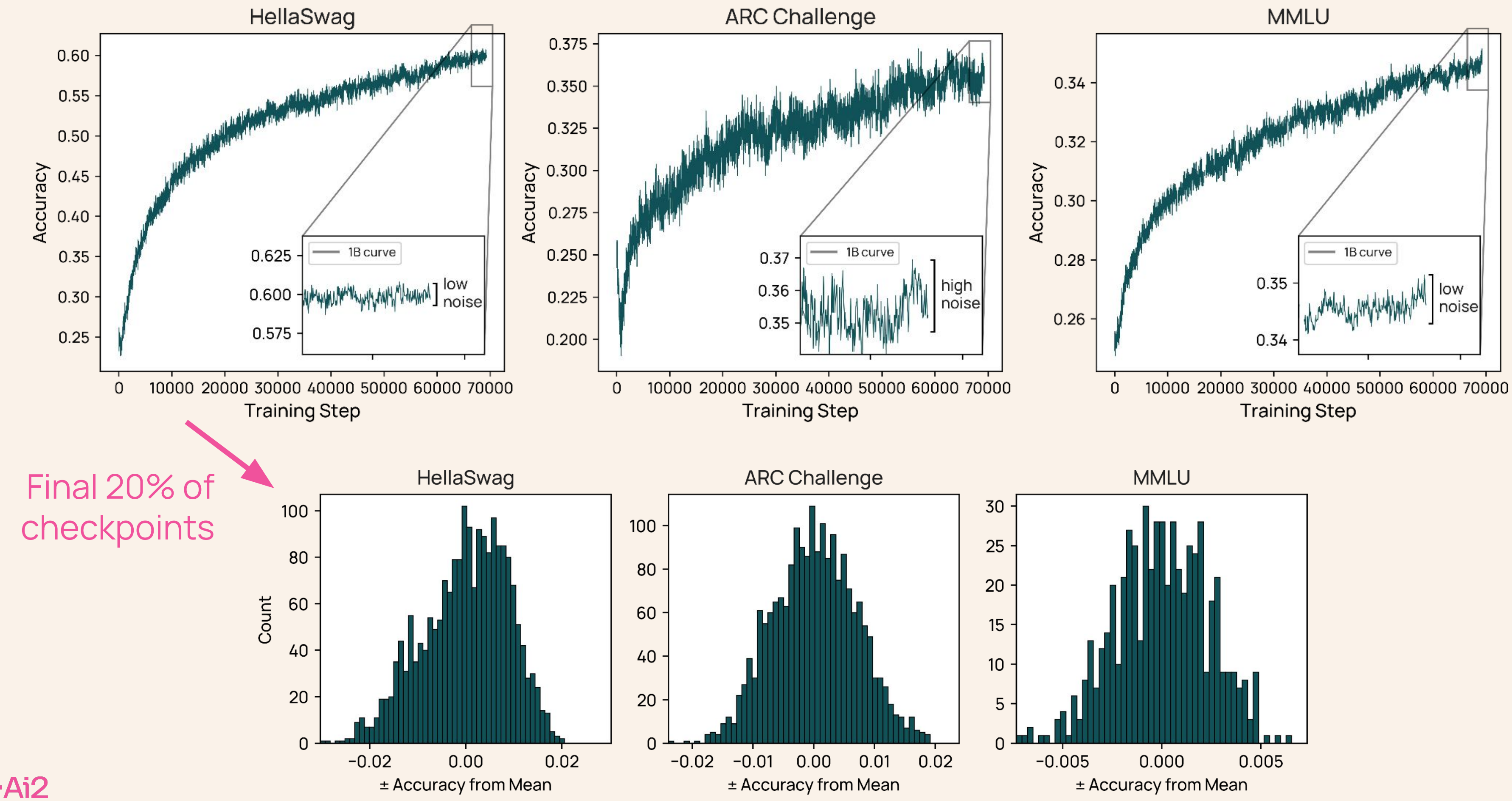




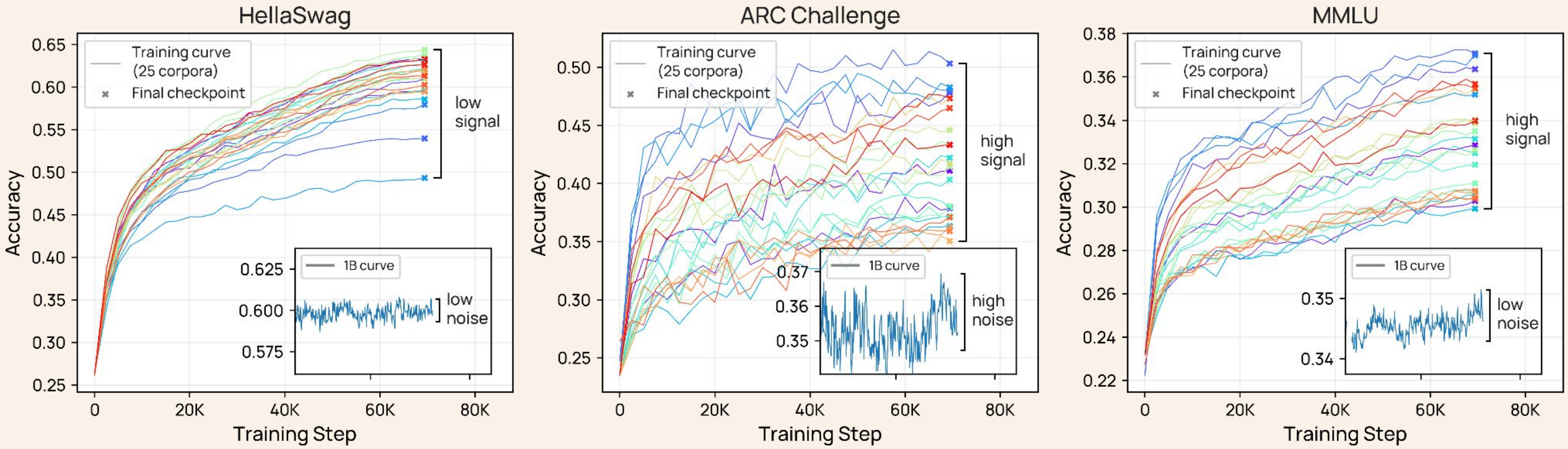
The Llama 3 Herd of Models (preprint, 2024)
Language models scale reliably with over-training and on downstream tasks (ICLR, 2025)
Are Emergent Abilities of Large Language Models a Mirage? (NeurIPS, 2023)

Ai2

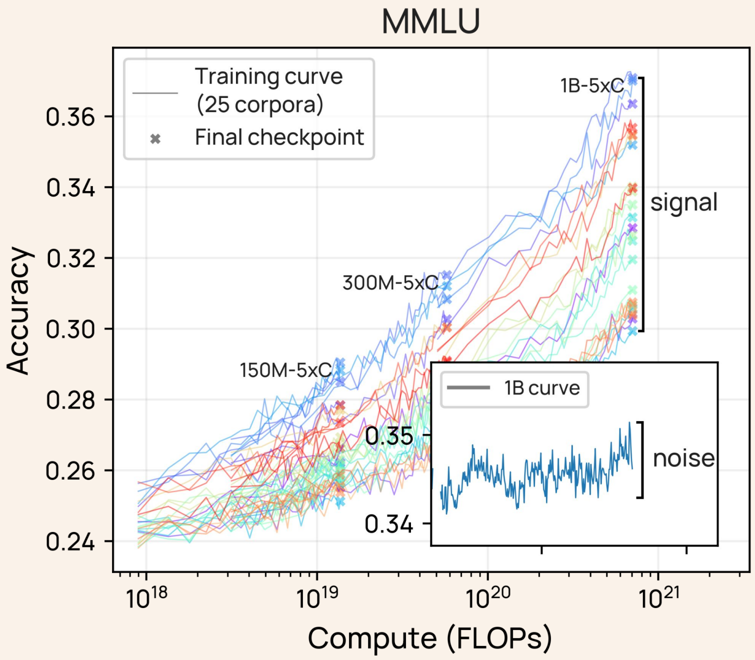
Why do so many predictions fail - but some don't?


HellaSwag
ARC Challenge
MMLU
Accuracy
Training Step
0 10000 20000 30000 40000 50000 60000 70000
0.60 0.55 0.50 0.45 0.40 0.35 0.30 0.25
0.625 0.600 0.575
1B curve
low noise
0.375 0.350 0.325 0.300 0.275 0.250 0.225 0.200
0.37 0.36 0.35
high noise
0.34 0.32 0.30 0.28 0.26
0.35 0.34
Final 20% of checkpoints
Count
± Accuracy from Mean
100 80 60 40 20 0
-0.02 0.00 0.02
-0.02 -0.01 0.00 0.01 0.02
30 25 20 15 10 5 0
-0.005 0.000 0.005

… but inter-checkpoint variance is not the whole story! We need to measure both signal and noise



Ai2



signal:

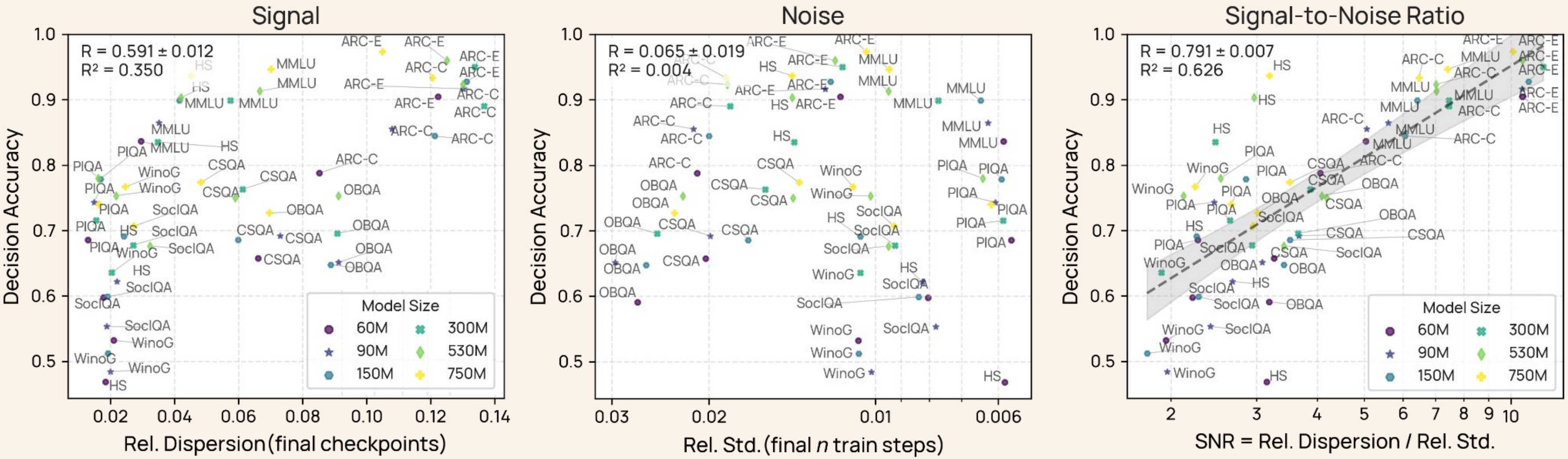
$$\text{Rel. Dispersion}(M) = \max_{j,k} |m_j - m_k| / \bar{m}$$

noise:

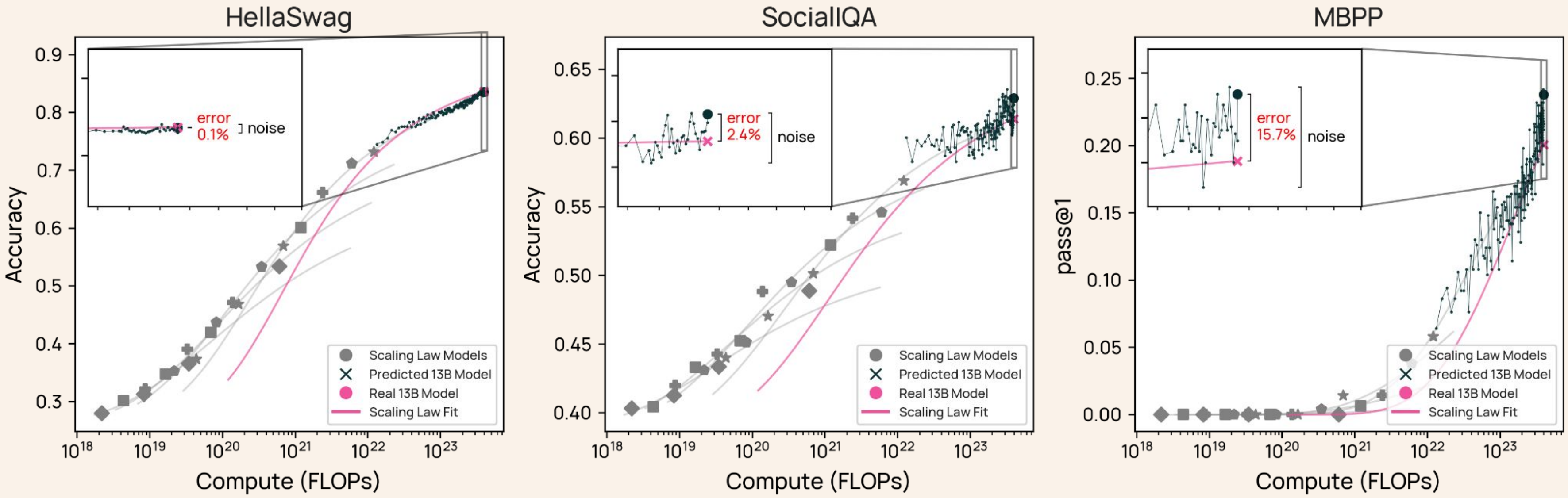
$$\text{Rel. Std.}(m) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (m_i - \bar{m})^2} / \bar{m}$$

$$\text{Signal-to-Noise Ratio} = \frac{\text{Rel. Dispersion(final train checkpoint)}}{\text{Rel. Std.(final } n \text{ train checkpoints)}}$$

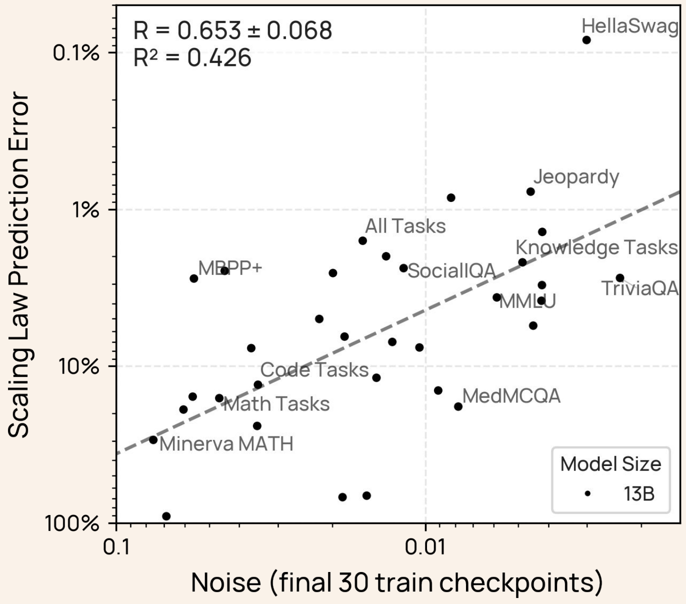
Only signal or noise alone do not explain rank agreement from small to large scale… … but the signal-to-noise ratio does!



Predicting task performance using scaling laws is sensitive to noise!



Predicting task performance using scaling laws is sensitive to noise!



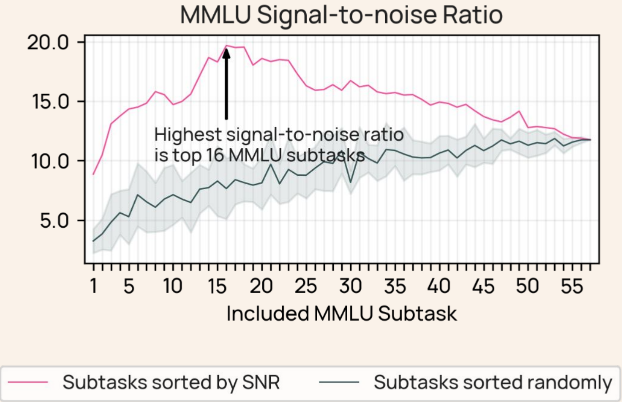
Ai2

We can use the signal-to-noise ratio to improve our benchmarks

➤ intervention 1: filter subtasks with high SNR

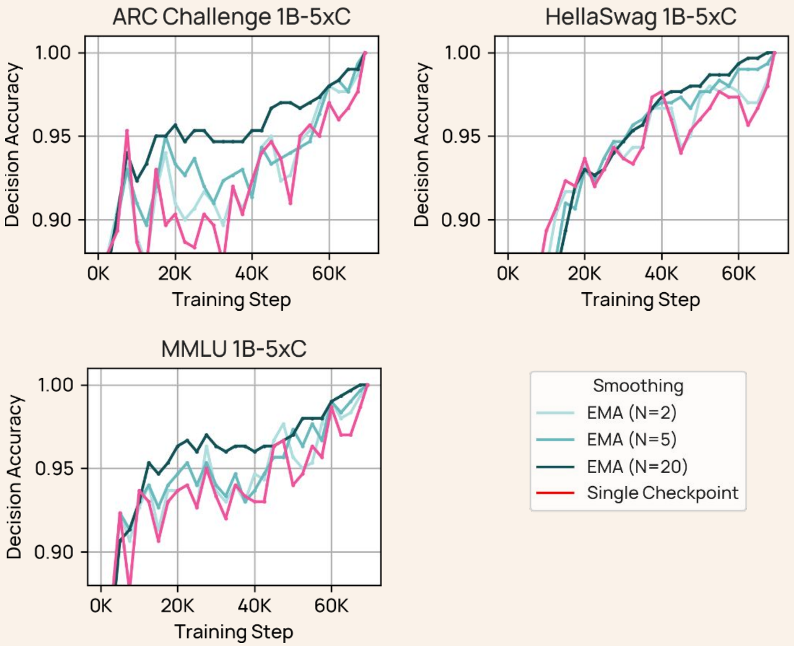
➤ intervention 2: smooth intermediate checkpoints

➤ intervention 3: select metrics with high SNR



We can use the signal-to-noise ratio to improve our benchmarks

- intervention 1: filter subtasks with high SNR
- **intervention 2: smooth intermediate checkpoints**
- intervention 3: select metrics with high SNR

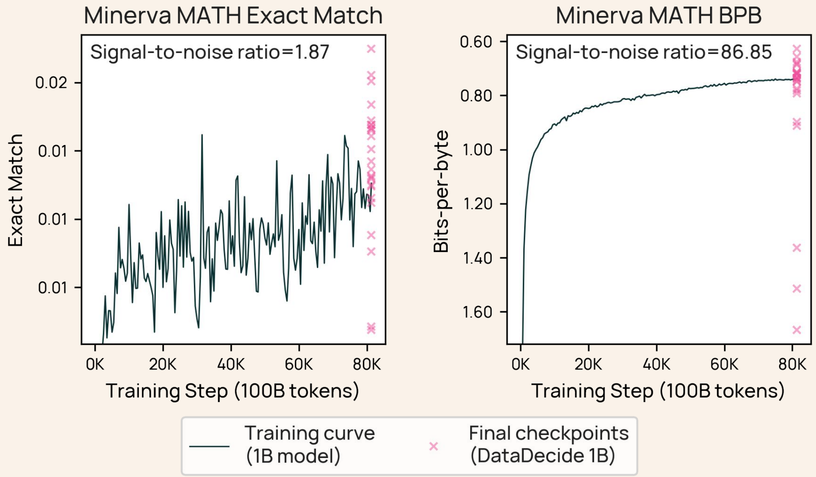


We can use the signal-to-noise ratio to improve our benchmarks

➤ intervention 1: filter subtasks with high SNR

➤ intervention 2: smooth intermediate checkpoints

➤ **intervention 3: select metrics with high SNR**



Thank you!

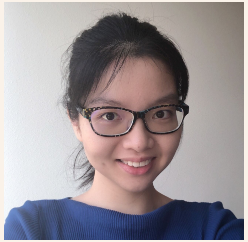
Learn more at our poster:

Wednesday, Dec 3 at 4:30 in Hall C,D,E.

Contact: `davidh@allenai.org`

NEURAL INFORMATION PROCESSING SYSTEMS







Ai2