

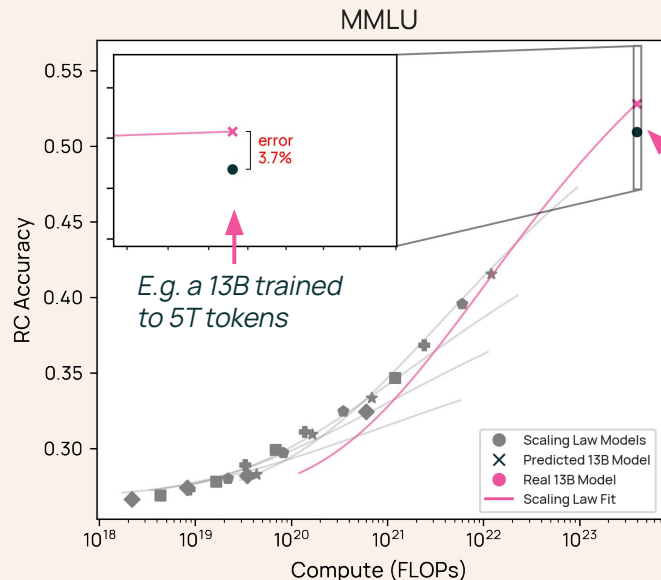
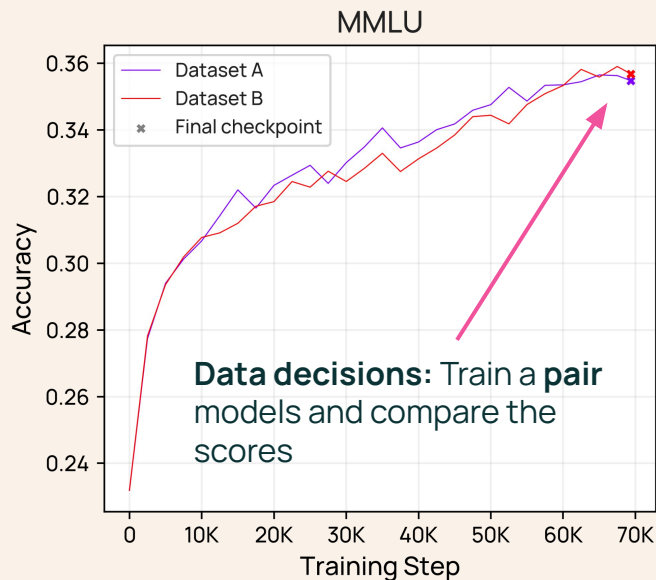
★ **Spotlight** poster on
Wednesday, Dec 3 at 4:30
★ in Hall C,D,E

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)

Downstream tasks are now core to building models ...

... some tasks are **useful!**

... some tasks **aren't predictable**

... some metrics **hide real capability**

Meta

The Llama 3 Herd of Models

Llama Team, AI @ Meta

A detailed contributor list can be found in the appendix of this paper.

Modern artificial intelligence (AI) systems are powered by foundation models. This paper presents a new set of foundation models, called Llama 3. It is a herd of language models that actively support multilinguality, coding, reasoning, and tool usage. Our largest model is a dense Transformer with 405B parameters and a context window of up to 128K tokens. This paper presents an extensive empirical evaluation of Llama 3. We find that Llama 3 delivers comparable quality to leading language models such as GPT-4 on a plethora of tasks. We publicly release Llama 3, including pre-trained and post-trained versions of the 405B parameter language model and our Llama Guard 3 model for input and output safety. The paper also presents the results of experiments in which we integrate image,

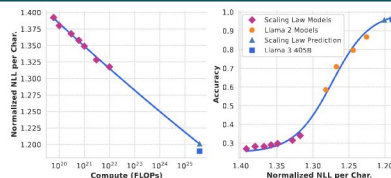


Figure 4: Scaling law forecast for ARC Challenge. *Left:* Normalized negative log-likelihood of the correct answer on the ARC Challenge benchmark as a function of pre-training FLOPs. *Right:* ARC Challenge benchmark accuracy as a function of the normalized negative log-likelihood of the correct answer. This analysis enables us to predict model performance on the ARC Challenge benchmark before pre-training commences. See text for details.

quality of the data we use for pre-training and post-training. These improvements include the development of more careful pre-processing and curation pipelines for pre-training data and the development of more rigorous quality assurance and filtering approaches for post-training data. We pre-train Llama 3 on a corpus of about 15T multilingual tokens, compared to 1.8T tokens for Llama 2.

- **Scale.** We train a model at far larger scale than previous Llama models: our flagship language model was pre-trained using 3.8×10^{16} FLOPs, about 50x more than the largest version of Llama 2. Specifically, we pre-trained a flagship model with 405B trainable parameters on 15.6T text volume. As expected per

Language models scale reliably with over-training and on downstream tasks

Samir Yitzhak Gadre^{1,2} Georgios Smyrnis³ Vaishal Shankar⁴
Suchin Gururangan³ Mitchell Wortman³ Rulin Shao³ Jean Mercat²
Alex Fang² Jeffrey Li² Sedrick Keh² Rui Xin² Marianna Nezhurina²
Igor Vasiljević² Justin Jitavee² Luca Soldaini³ Alexandre G. Dimakis³
Gabriel Ilharco² Pang Wei Koh^{3,5} Shuran Song² Thomas Kollar²
Yair Carmon² Achal Dave² Reinhard Heckel^{1,6} Niklas Muennighoff^{1,6} Ludwig Schmidt^{1,6}

Jun 2024

May 2023

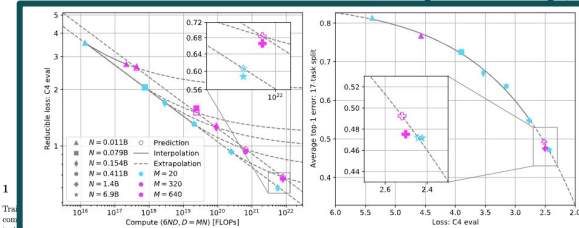


Figure 1: Reliable scaling with over-training and on downstream error prediction. *(left)* We fit a scaling law for model validation loss, parameterized by (i) a token multiplier $M = N/D$, which is the ratio of training tokens D to parameters N and (ii) the compute C in FLOPs used to train a model, approximated by $C = 6ND$. Larger values of M specify more over-training. We are able to extrapolate, in both N and M , the validation performance of models requiring more than 300x the training compute used to construct the scaling law. *(right)* We also fit a scaling law to predict average downstream top-1 error as a function of validation loss. We find that fitting scaling laws for downstream error benefits from using more expensive models when compared to fitting for loss prediction. We predict the average error over 17 downstream tasks for models trained with over 20x the compute. For this figure, we train all models on RedPajama [112].

Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer, Brando Miranda, and Sami Koyejo
Computer Science, Stanford University

Abstract

Recent work claims that large language models display emergent abilities, such

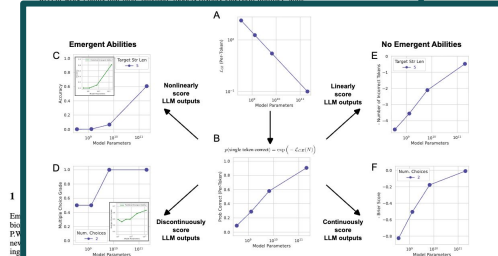


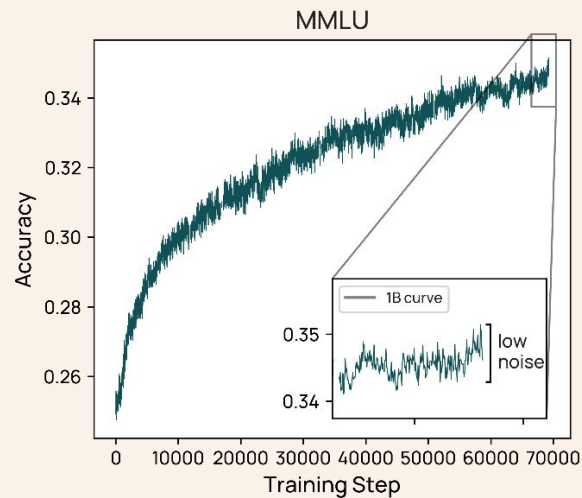
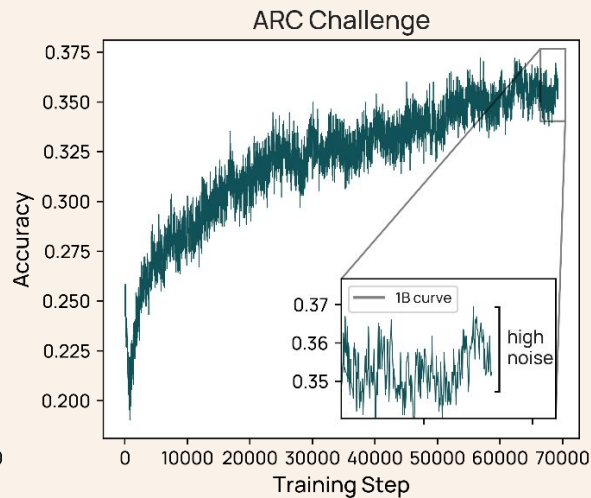
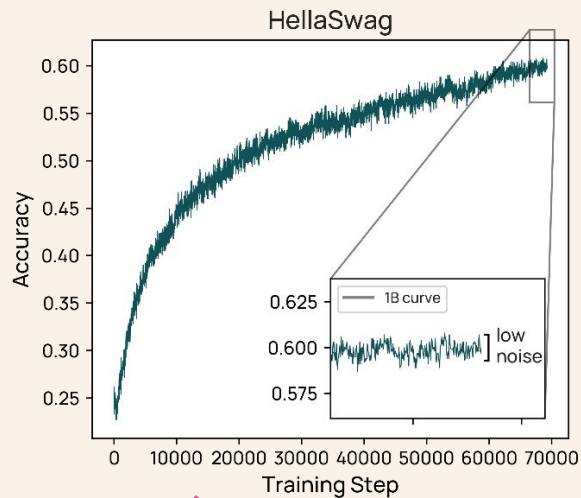
Figure 2: Emergent abilities of large language models are created by the researcher's chosen metrics, not unpredictable changes in model behavior with scale. (A) Suppose the per-token cross-entropy loss decreases monotonically with model scale, e.g., LCE scales as a power law. (B) The per-token probability of selecting the correct token asymptotes towards 1. (C) If the researcher scores models' outputs using a nonlinear metric such as Accuracy (which requires a sequence of tokens to all be correct), the metric choice nonlinearly scales performance, causing performance to change sharply and unpredictably in a manner that qualitatively matches published emergent abilities (inset). (D) If the researcher instead scores models' outputs using a discontinuous metric such as Multiple Choice Grade (akin to a step function), the metric choice discontinuously scales performance, again causing performance to change sharply and unpredictably. (E) Changing from a nonlinear metric to a linear metric such as Token Edit Distance, scaling shows smooth, continuous and predictable improvements, ablating the emergent ability. (F) Changing from a discontinuous metric to a continuous metric such as Brier Score again reveals smooth, continuous and predictable improvements in task performance. Consequently, emergent abilities are created by the researcher's choice of metrics, not fundamental changes in model family behavior on specific tasks with scale.

The Llama 3 Herd of Models (preprint, 2024)

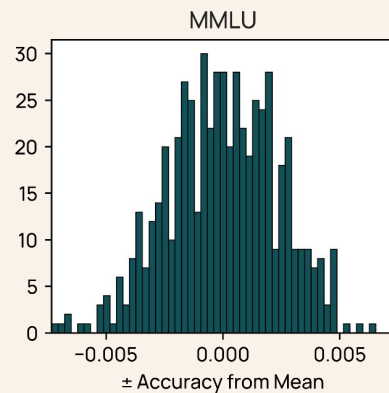
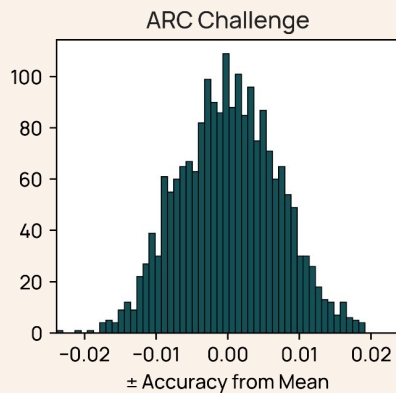
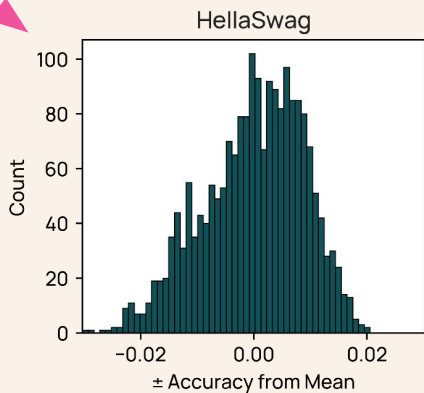
Language models scale reliably with over-training and on downstream tasks (ICLR, 2025)

Are Emergent Abilities of Large Language Models a Mirage? (NeurIPS, 2023)

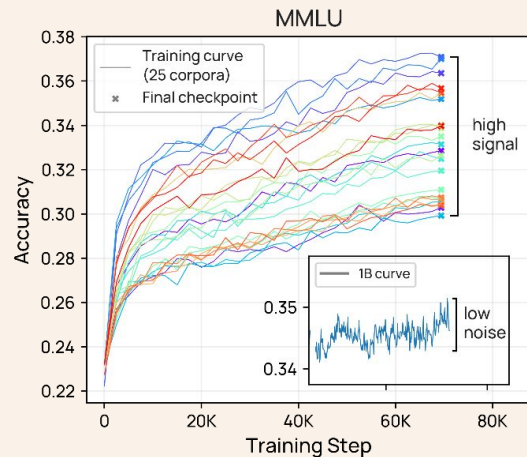
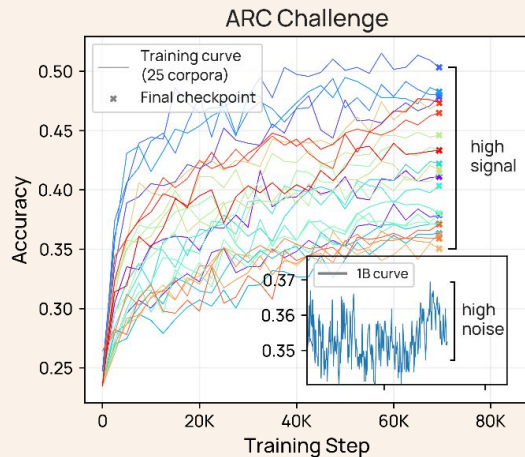
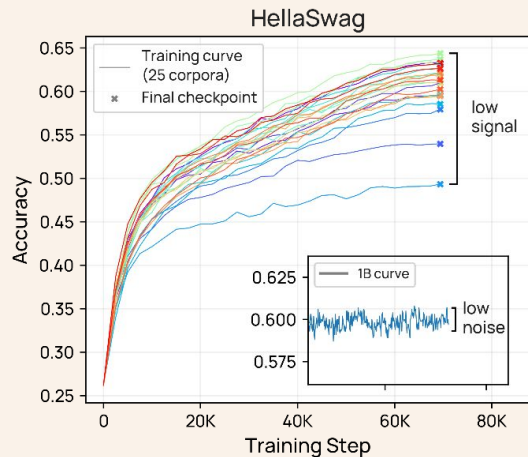
Why do so many predictions
fail - but some don't?

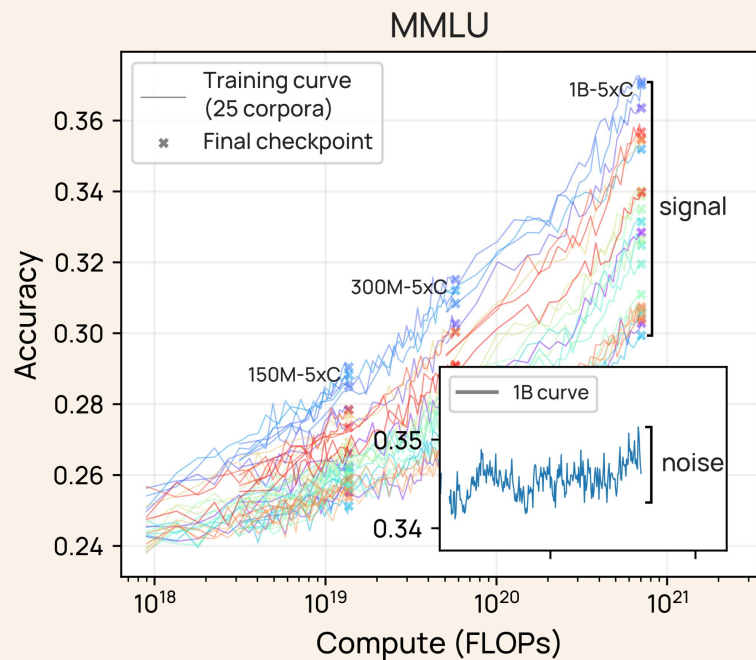


Final 20% of checkpoints



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





signal:

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

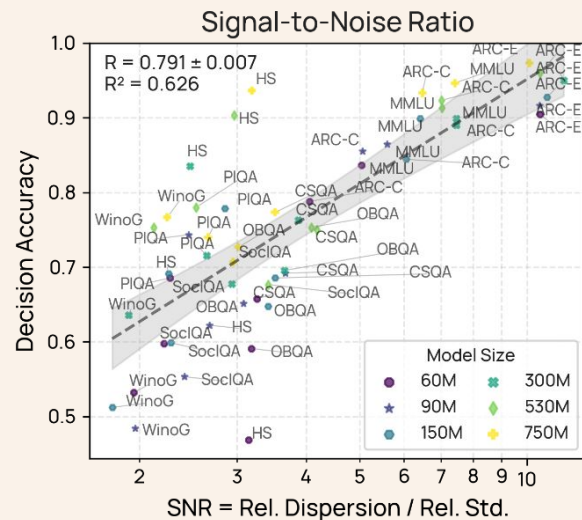
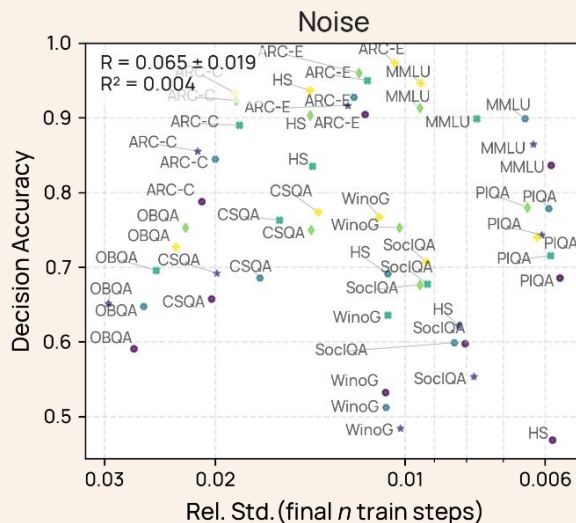
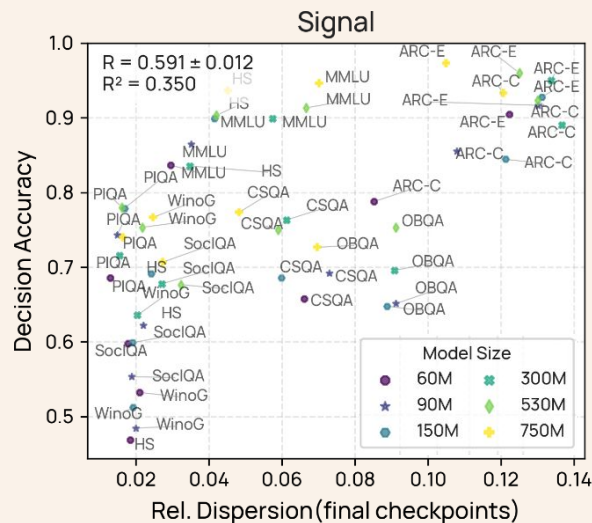
noise:

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

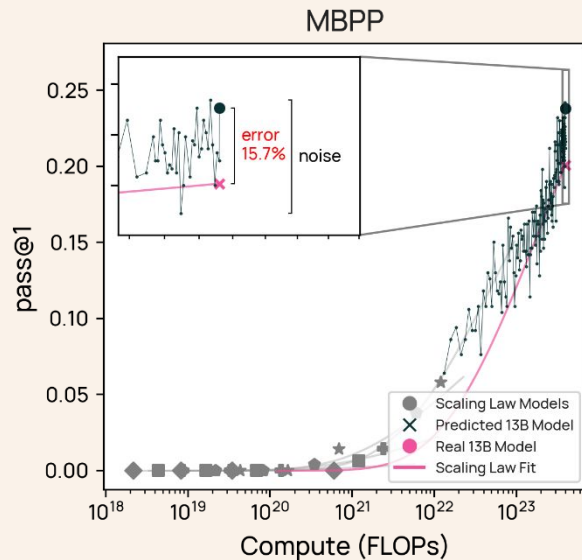
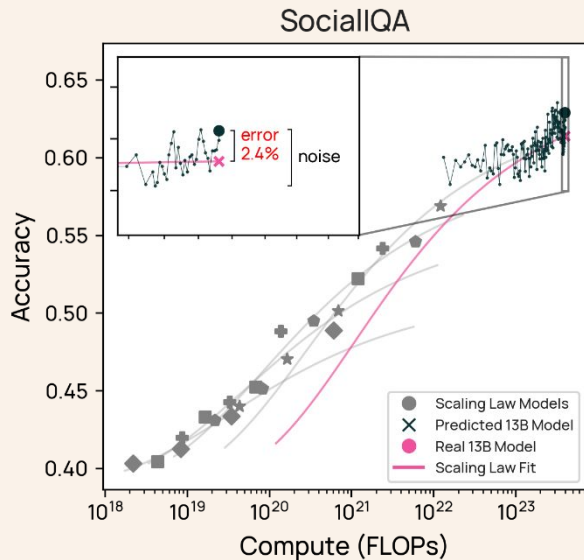
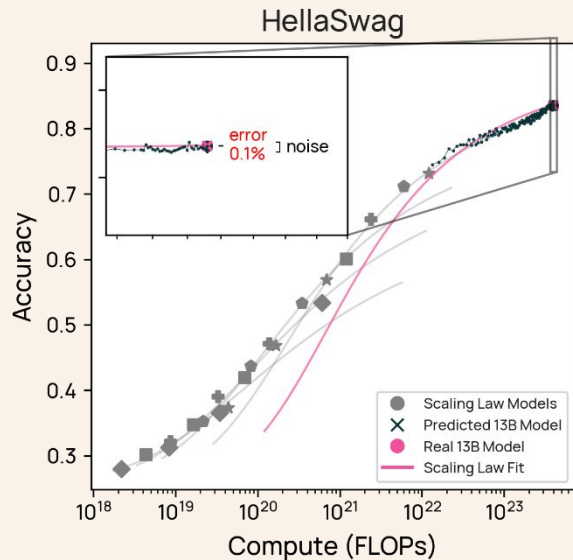
Signal-to-Noise Ratio =

$$\frac{\text{Rel. Dispersion}(\text{final train checkpoint})}{\text{Rel. Std.}(\text{final } n \text{ train checkpoints})}$$

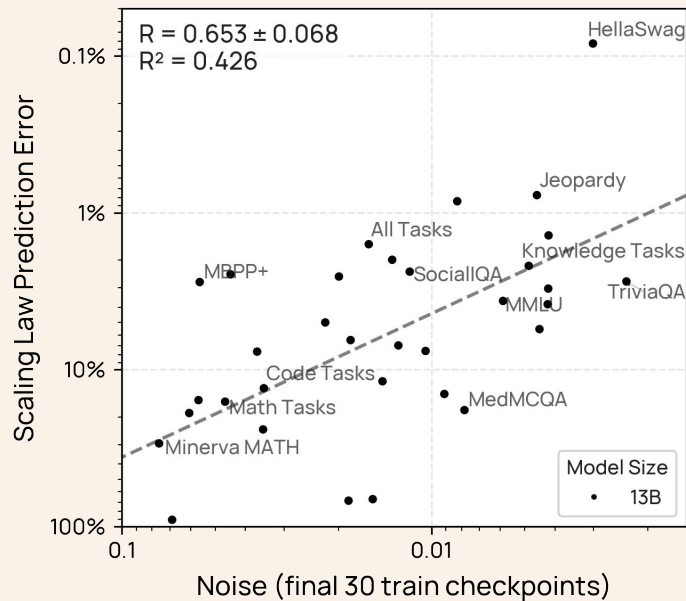
Only **signal** or **noise** alone do not explain rank agreement from small to large scale... ... but the **signal-to-noise ratio** does!



Predicting task performance using scaling laws is sensitive to noise!

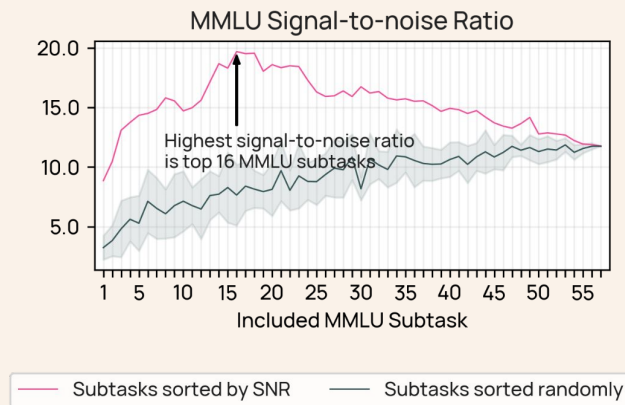


Predicting task performance using scaling laws is sensitive to noise!



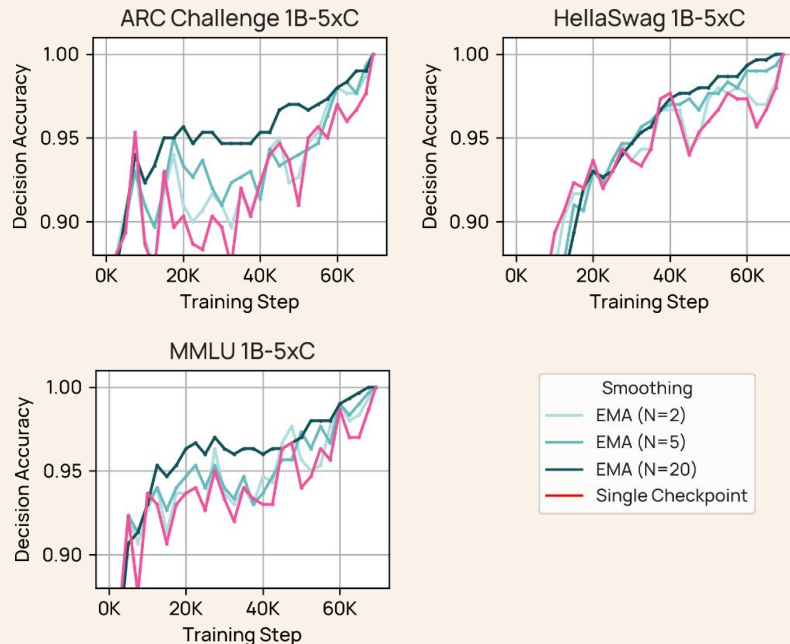
We can use the signal-to-noise ratio to **improve our benchmarks**

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



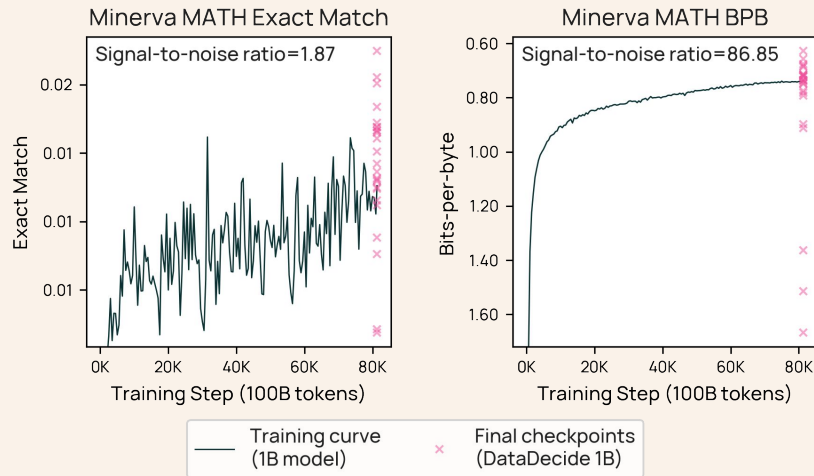
We can use the signal-to-noise ratio to **improve our benchmarks**

- 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

