Great Models Think Alike and this Undermines *AI Oversight*

Shashwat Goel[E,M], Joschka Strüber[T,U], Ilze Amanda Auzina[T,U], Karuna K C[I],
Ponnurangam K[I], Douwe Kiela[C,S], Ameya Prabhu[T,U], Matthias Bethge[T,U], Jonas Geiping[E,M,T]

ICML International Conference On Machine Learning

E ellis Institute Tübingen · M Max Planck Institute for Intelligent Systems · S Stanford University · U Eberhard Karls Universität Tübingen · C contextual.ai · T Tübingen AI Center · I International Institute of Information Technology Hyderabad

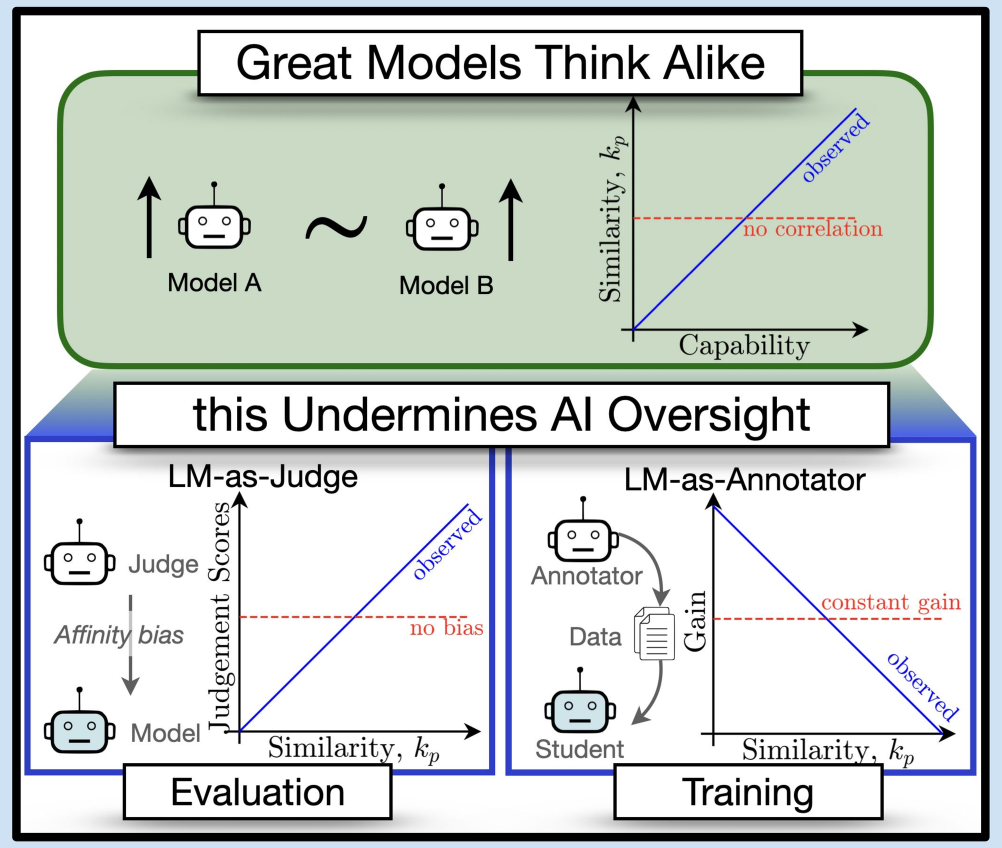
⬇ Only 2 *MINUTES* ⏰ ? Read Here ⬇

AI Oversight = Using models to ***evaluate*** and annotate ***training*** data for other models

Our FINDINGS

As capabilities increase we defer more to **AI oversight…**

Training LLM judges show *affinity bias* - they favor similar models



Models make similar mistakes as capabilities increase!

Evaluation Complementary knowledge explains gains in weak-to-strong

Novel Model SIMILARITY Metric

⟪*Chance Adjusted Probabilistic Agreement (CAPA)*⟫

Similarity Metric	Adjusts for Accuracy	Distinguishes different mistakes	Incorporates Probabilities
%Flips	❌	❌	❌
Cohen's κ	❌	✅	❌
%Agreement	❌	✅	❌
Error Cons.	✅	❌	❌
Pearson's ρ	✅	❌	❌
KL, JS Div	❌	✅	✅
CAPA $(\kappa)_p$	✅	✅	✅

💡 Similar models make similar predictions

⚠️ Models can have similar predictions by virtue of high accuracy.

Use probability distrib. over predictions instead of sampling

Think two models have similar behavior? 🤨
or Some interventions have complementary benefits? ☯️
or Using multiple models or judges together will help? 🤖👾🦾

You can now ***quantify*** similarity! — **pip install lm-sim**

WHAT WE *DO*

Effect of Similarity on LLM-as-a-Judge

→ Evaluate on **MMLU Pro - 14 domains**
→ Filter questions for free-form evaluation
→ Use **LLM-as-a-judge** to rate free-form answers
→ Pairs across 9 judges and 39 judged models

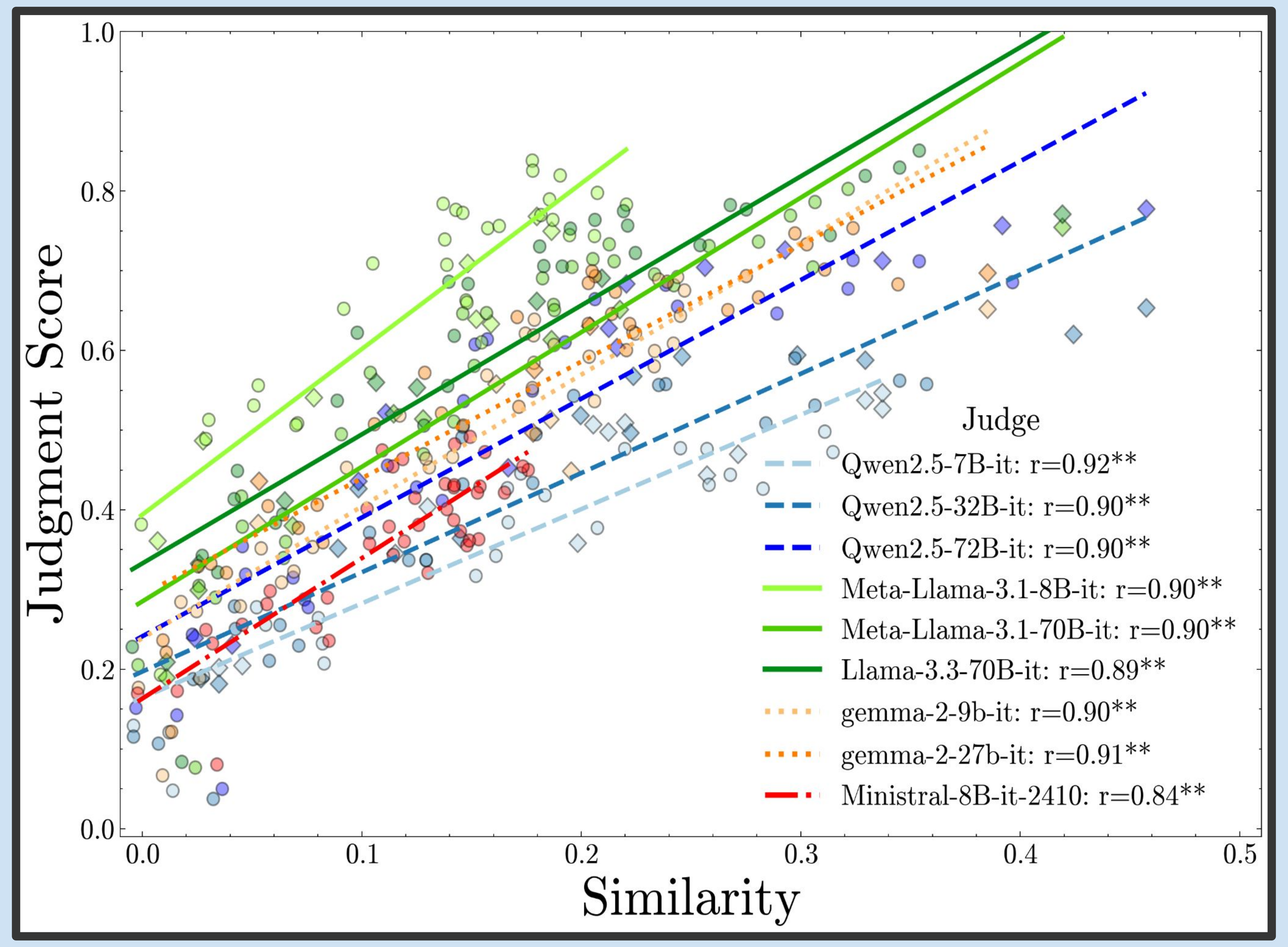
Effect of Similarity on Weak-to-Strong Training

→ **OpenAI Weak-to-strong generalization** setup
→ **Models**: Weak 1-3B, Strong 7-9B parameters
→ Studied 12 model pairs on 15 NLP tasks

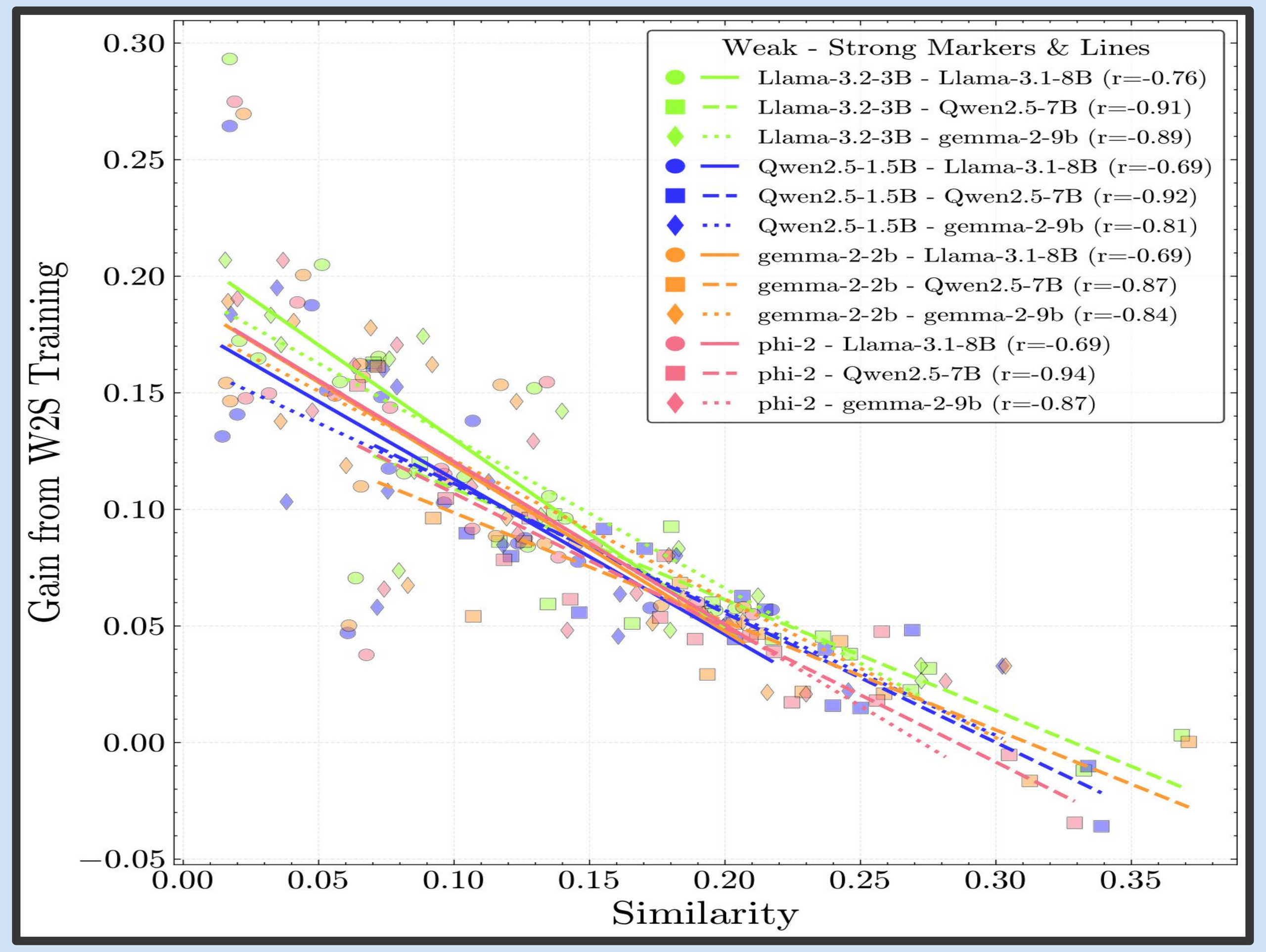
Effect of Improving Capability on Similarity

→ **130 models** from 🤗 OpenLLM Leaderboard
→ Datasets: **MMLU Pro** & **Big Bench Hard**
→ Legend: Color = family, Size = #Params

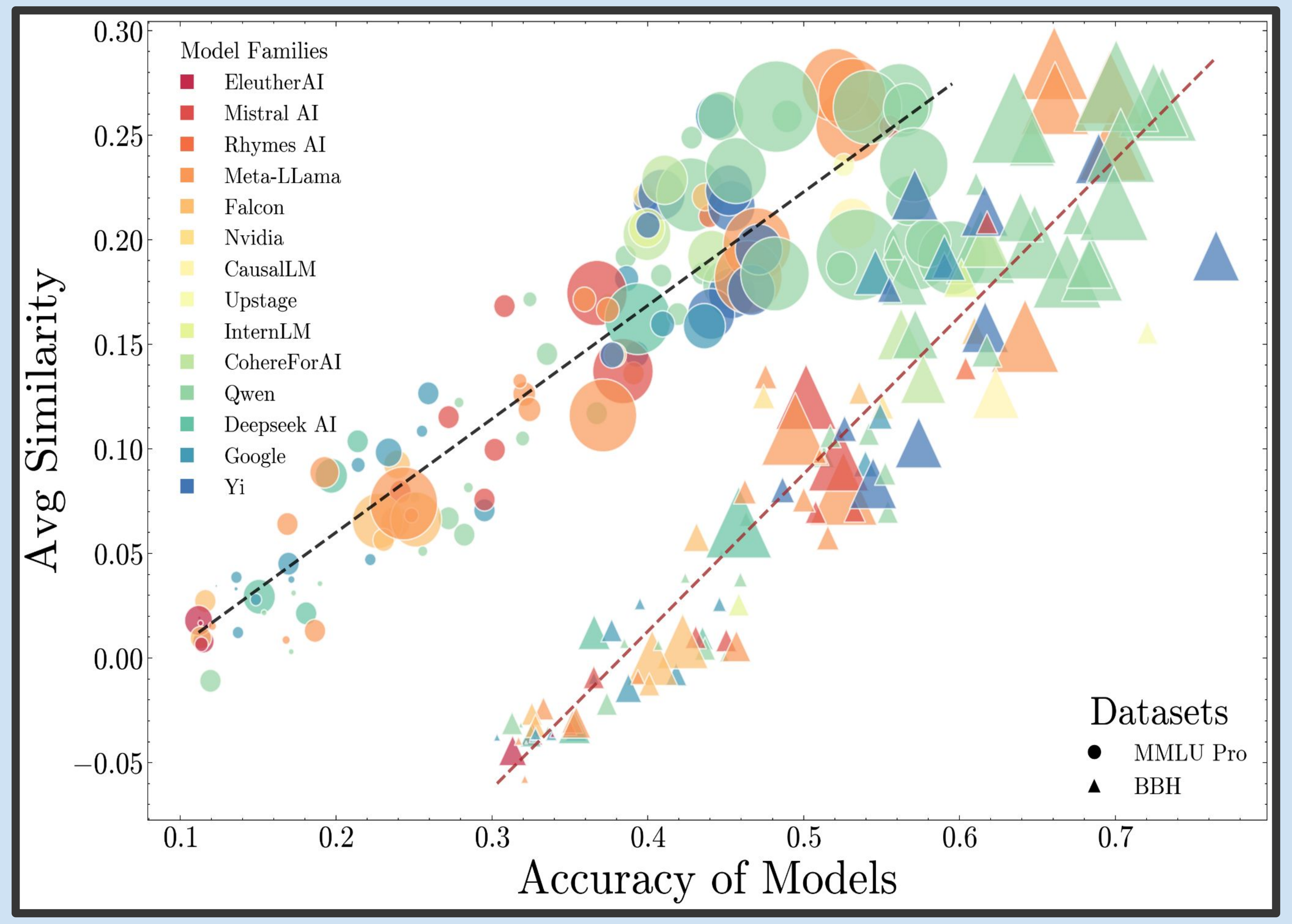
WHAT WE *FIND*



Affinity Bias: Judgement scores increase with similarity, even when controlling for true accuracy



Training on LM annotations benefits from complementary knowledge



With increasing capabilities, model errors are becoming more correlated

How to measure Similarity

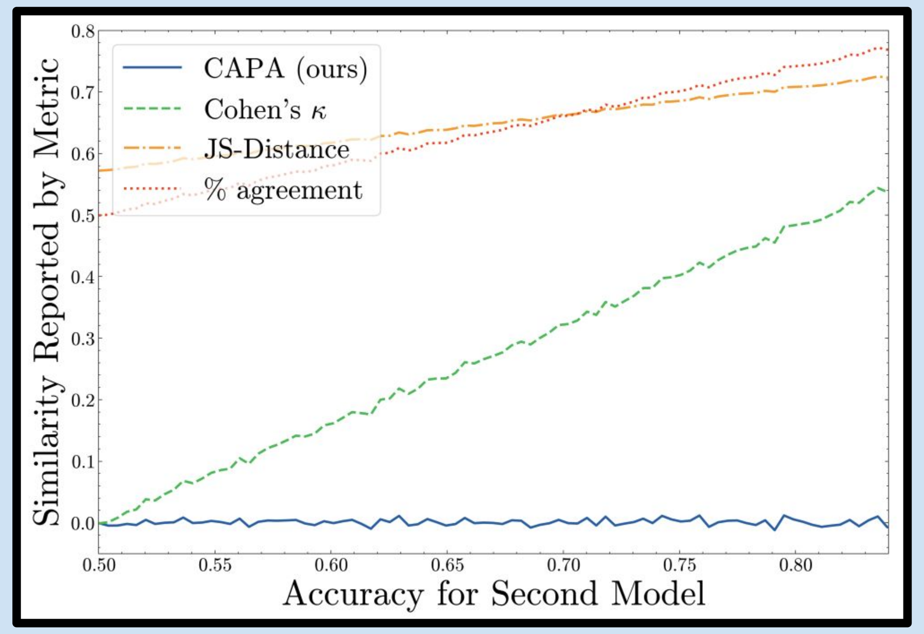
$$c^p_{obs} = \frac{1}{|D|}\sum_{x\in D}\sum_{o_i\in O(x)} p_1(o_i)\cdot p_2(o_i)$$

$$\kappa_p = \frac{c^p_{obs} - c^p_{exp}}{1 - c^p_{exp}}$$

$$c^p_{exp} = \underbrace{\overline{p_1}\cdot\overline{p_2}}_{\text{chance agreement on correct option}} + \underbrace{(1-\overline{p_1})\cdot(1-\overline{p_2})\cdot\frac{1}{|D|}\sum_{x\in D}\frac{1}{|O(x)|-1}}_{\text{chance agreement on incorrect option}}$$

c^p_{obs} - ***Observed Agreement***
Probability of agreement if the model's predictions were sampled based on the ***observed likelihoods*** assigned over options

c^p_{exp} - ***Expected Agreement***
To account for higher accuracies inflating observed agreement, ***normalize*** by the agreement expected from two independent models.



Metric comparison for ***independent models*** with **uncorrelated predictions**. CAPA correctly reports **0** ***similarity*** when models have uncorrelated errors.

Paper, Data, Code and Demo!

pip install lm-sim