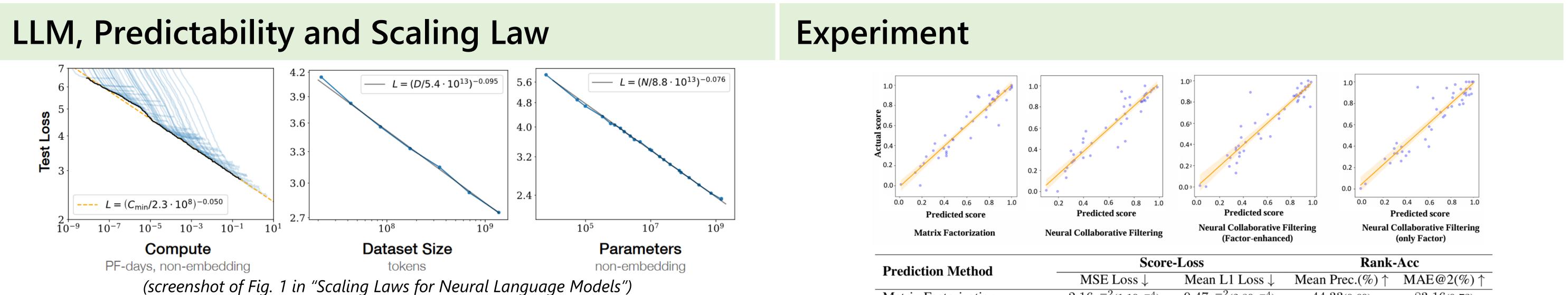




Collaborative Performance Prediction for Largelanguage Model Evaluation

Qiyuan Zhang, Fuyuan Lyu^{*}, Xue Liu and Chen Ma^{*}



Power-law relationship $\log(L_m) \approx \omega_f \log(C_m) + b_f$,

> High Cost:

- (a) Train: repeated training model in a family
- (b) Test: e.g. Chain-of-Thought (CoT)
- > Lacking consideration of non-computational factors

E.g. Data Quality, #shots

> Ignoring relationships among models and tasks. Prediction **limited to:** (a) one model family and (b) one metric

Beyond Scaling Law

ELM Leade	ws how the various model						ect a group: Core scenarios
Accuracy Calibration	Robustness Fairness Mean win rate	Efficiency General inform	BoolQ - EM	NarrativeQAE1	NaturalQuestions.(closed)	ed)E1. 🗘 👘 NaturalQuestions (open)E1. 🗘	QuACE1
Llama 2 (70B)	0.944 ଙ	0.582	0.886	0.77	0.458	0.674	0.484
LLaMA (65B)	0.908 2	0.584	0.871	0.755	0.431	0.672	0.401
text-davinci-002	0.905 虑	0.568	0.877	0.727	0.383	0.713	0.445
Mistral v0.1 (7B)	0.884 🖻	0.572	0.874	0.716	0.365	0.687	0.423
Cohere Command beta (52.4B)	0.874 @	0.452	0.856	0.752	0.372	0.76	0.432
text-davinci-003	0.872 虑	0.569	0.881	0.727	0.406	0.77	0.525
Jurassic-2 Jumbo (178B)	0.824 @	0.48	0.829	0.733	0.385	0.669	0.435
Llama 2 (13B)	0.823 🕫	0.507	0.811	0.744	0.376	0.637	0.424
TNLG v2 (530B)	0.787 🗭	0.469	0.809	0.722	0.384	0.642	0.39
gpt-3.5-turbo-0613	0.783 🖻	0.391	0.87	0.625	0.348	0.675	0.485
LLaMA (30B)	0.781 🗈	0.531	0.861	0.752	0.408	0.666	0.39

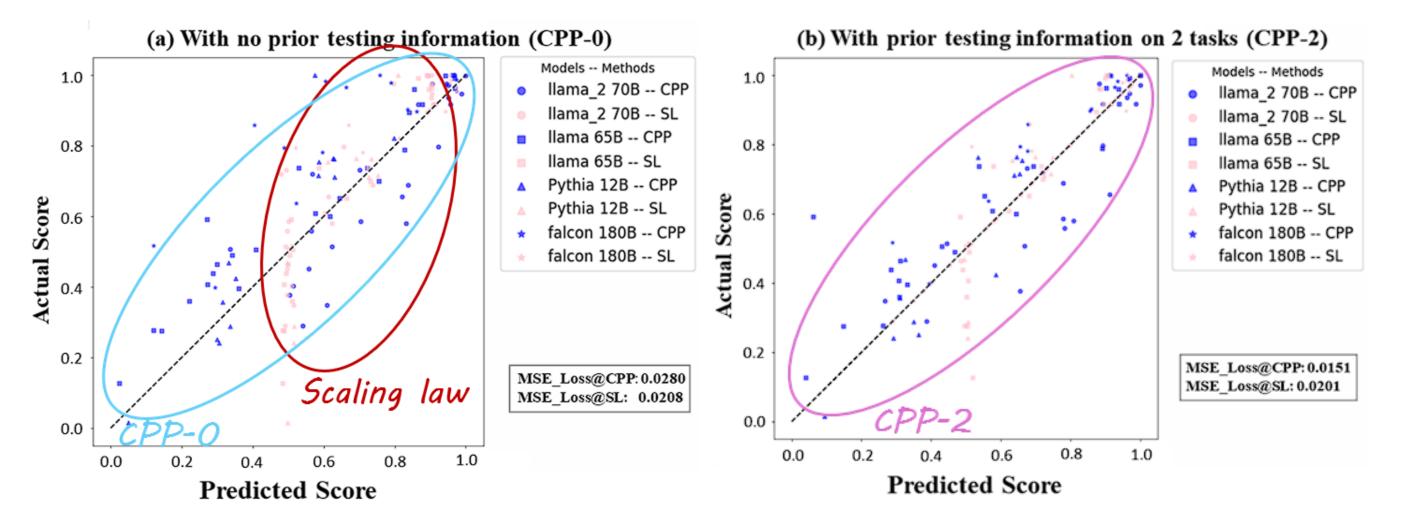
HELM Core Leaderboard 68 models, 16 tasks, 82.5% sparsity Solved as a Matrix Completion problem?

Matrix Factorization	Neural Collaborative Filtering	Neural Collaborative F (Factor-enhance)	-	ng Neural Collaborative Filtering (only Factor)	
Prediction Method	Score-Loss		Rank-Acc		
Prediction Method	MSE Loss \downarrow	Mean L1 Loss ↓	Mean Prec.(%) ↑	MAE@2(%) ↑	
Matrix Factorization	$2.16e^{-2}(1.19e^{-4})$	$9.47e^{-2}(2.89e^{-4})$	44.33(0.69)	83.16(0.73)	
Neural Collaborative Filtering	$1.58e^{-2}(4.22e^{-5})$	$8.94e^{-2}(3.10e^{-4})$	41.76(1.22)	84.98 (0.42)	
+ Factor Enhanced	$1.25e^{-2}(3.35e^{-6})$	$7.88e^{-2}(6.31e^{-5})$	45.45 (0.33)	84.54(0.27)	
Only Factor	$1.75e^{-2}(2.07e^{-5})$	$8.57e^{-2}(1.48e^{-4})$	33.47(0.12)	84.08(0.37)	

Observations:

- > Collaborative Performance Prediction is feasible Predicted Score \approx Gold Score Further Improvement Through
- a) Complex Model: NCF > MF b) Descriptive Factors: NCF(Factor Enhanced) > NCF
- > Support Predictions based only on descriptive factors

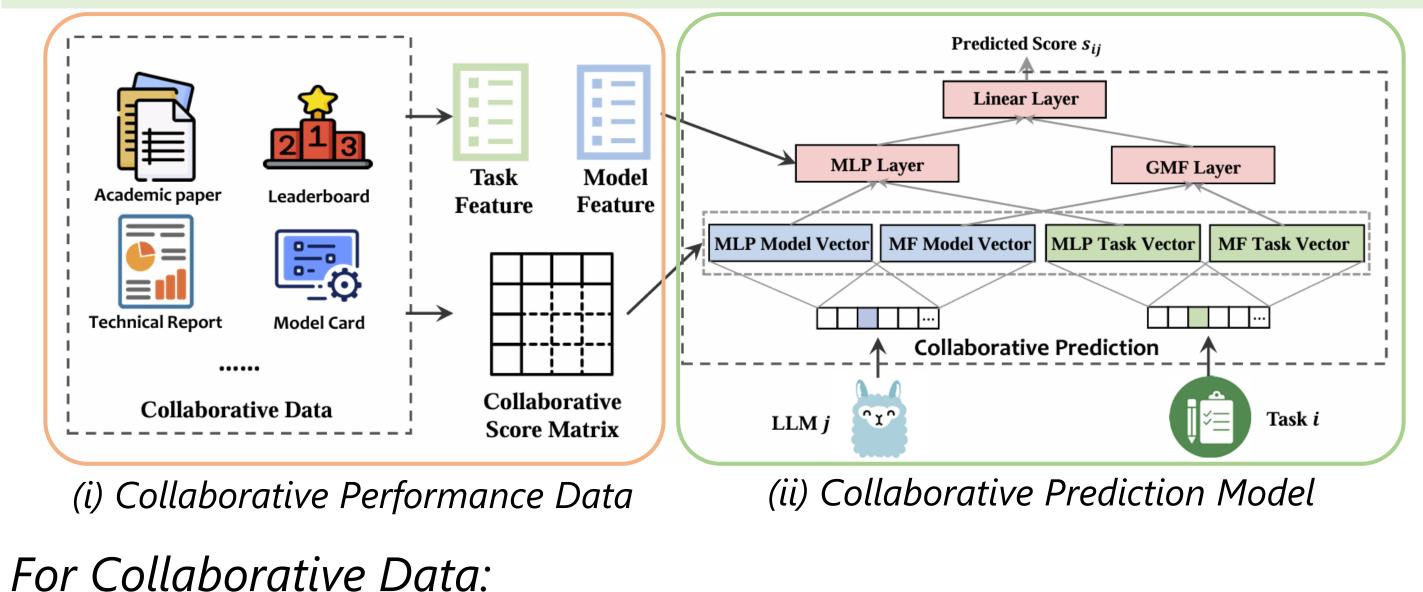
Generalization on New Model and Task





 \succ Matrix Factorization (MF) with Latent Vector = 10 \succ MF can accurately predict the missing scores with low error

Collaborative Performance Prediction (CPP)



Generalization on New Model:

CPP demonstrates greater adaptability than SL

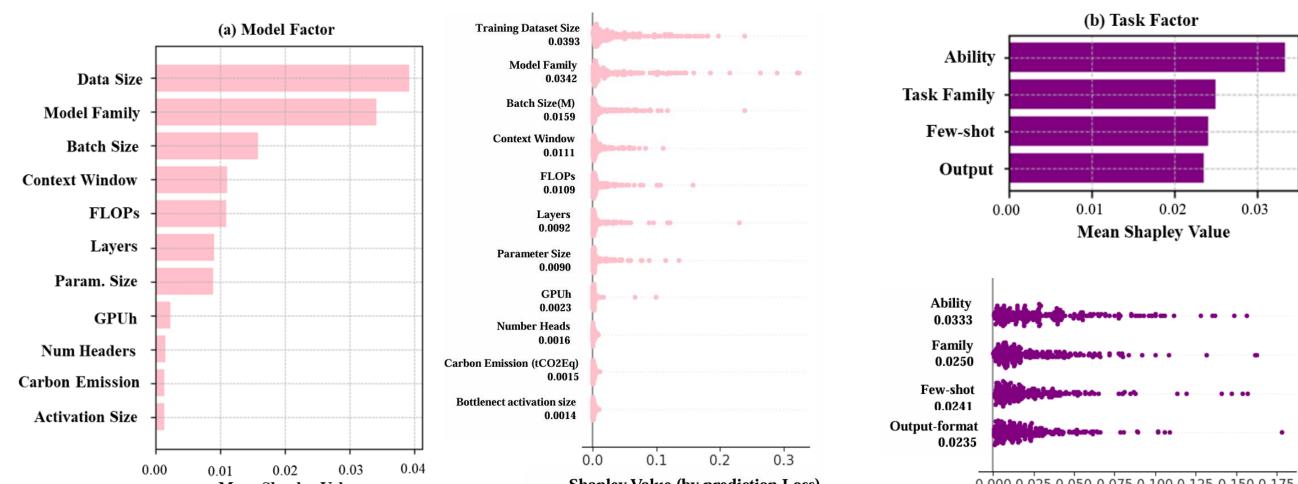
> Conducting Evaluation on a few tasks can improve Predictability

Models	BoolQ(0-shot)	BIG-bench hard(3-shot)	HellaSwag(10-shot)	HumanEval(pass@1)
CPP-T0	0.02201	0.07103	0.03414	0.1244
CPP-T2	0.0182	0.00725	0.02506	0.0763

Generalization on New Task:

> Conducting Evaluation on a few models can improve Predictability

Factor Importance Analysis

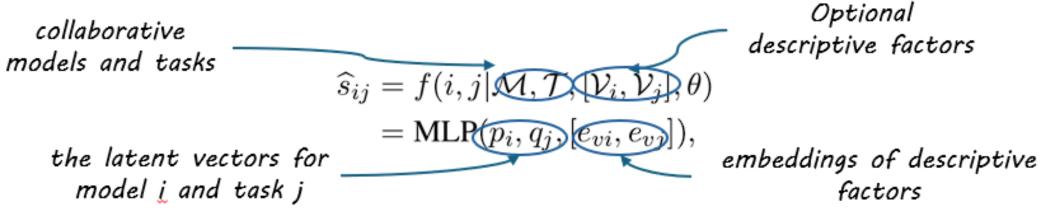


Existing Leaderboard e.g. HELM, OpenLLM, Compass Custom Leaderboard

3 Leaderboard Collect the collaborative 55 Paper/Technical Report performances 31 Model Card

For Collaborative Prediction Model: Matrix Factorization (MF)





Sparsity < 15%

Sparsity = 44%#Models = 72#Tasks = 2212 Model Factors 4 Task Factors

Mean Shapley Value

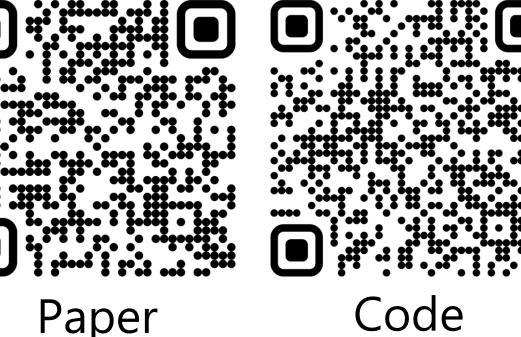
Shapley Value (by prediction Loss)

0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 **Shapley Value (by prediction Loss)**

Factor Analysis via Shapley Value: > Non-computational model factor and task factor are also vital

Summary & Future Work

- Predictability beyond Scaling Law
- Relationship Research among Models and Tasks: collaborative research via
 - open-source design factors
- Efficient Evaluation with Predictability



Code