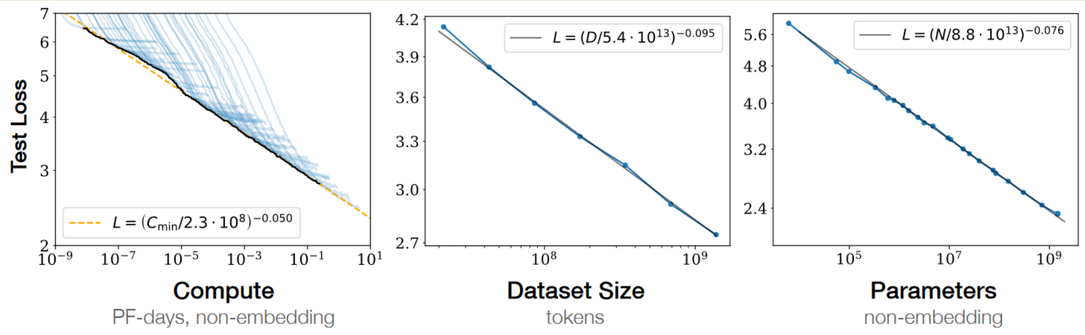
Collaborative Performance Prediction for Large-language Model Evaluation

Qiyuan Zhang, Fuyuan Lyu*, Xue Liu and Chen Ma*

LLM, Predictability and Scaling Law

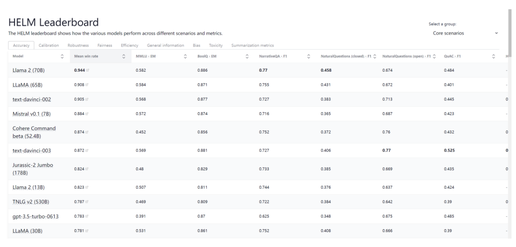


(screenshot of Fig. 1 in "Scaling Laws for Neural Language Models")

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

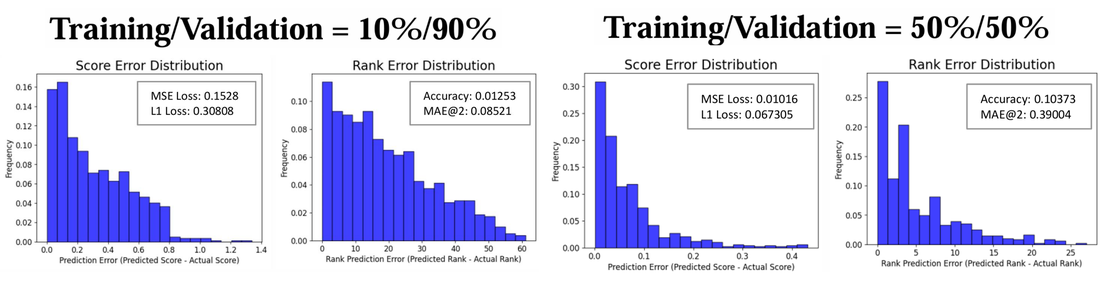
Beyond Scaling Law



HELM Core Leaderboard

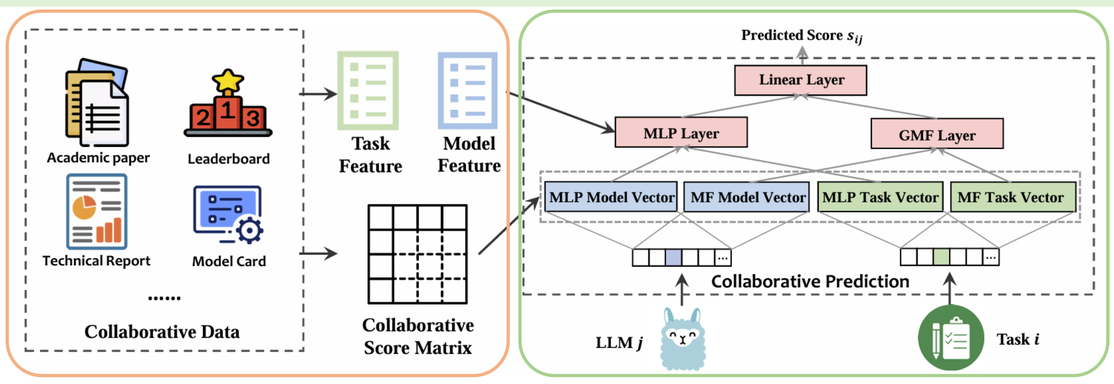
68 models, 16 tasks, 82.5% sparsity

Solved as a Matrix Completion problem?



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

Collaborative Performance Prediction (CPP)



(i) Collaborative Performance Data *(ii) Collaborative Prediction Model*

For Collaborative Data:

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

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

Sparsity < 15%

Sparsity = 44%
#Models = 72
#Tasks = 22
12 Model Factors
4 Task Factors

For Collaborative Prediction Model:

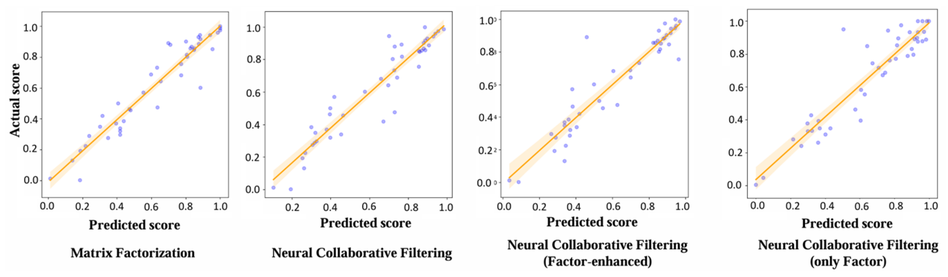
- Matrix Factorization (MF)
- Neural Collaborative Filtering (NCF)

collaborative models and tasks — *Optional descriptive factors*

$$\hat{s}_{ij} = f(i, j | \mathcal{M}, \mathcal{T}, [\mathcal{V}_i, \mathcal{V}_j], \theta) = \text{MLP}(p_i, q_j, [e_{vi}, e_{vj}]),$$

the latent vectors for model i and task j — *embeddings of descriptive factors*

Experiment

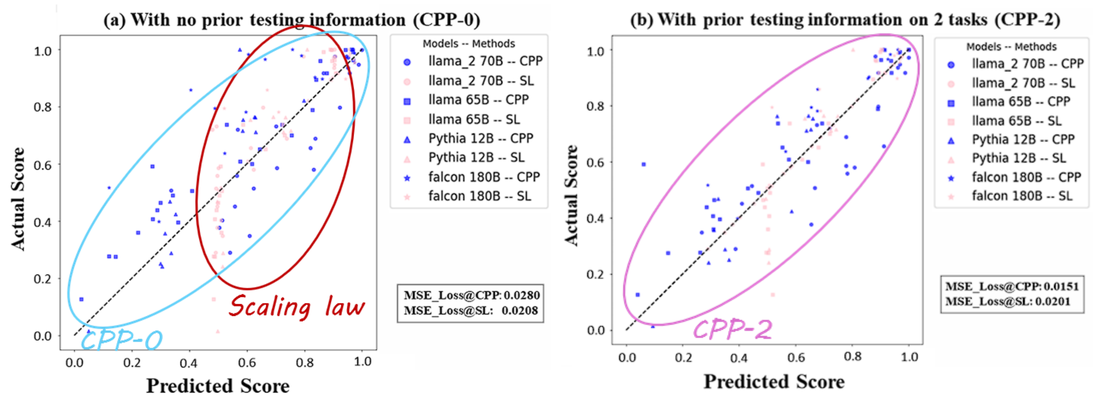


Prediction Method	Score-Loss		Rank-Acc	
	MSE Loss ↓	Mean L1 Loss ↓	Mean Prec.(%) ↑	MAE@2(%) ↑
Matrix Factorization	$2.16e^{-2}(1.19e^{-4})$	$9.47e^{-2}(2.89e^{-4})$	44.33(0.69)	83.16(0.73)
Neural Collaborative Filtering	$1.58e^{-2}(4.22e^{-5})$	$8.94e^{-2}(3.10e^{-4})$	41.76(1.22)	**84.98(0.42)**
+ Factor Enhanced	$\mathbf{1.25e^{-2}(3.35e^{-6})}$	$\mathbf{7.88e^{-2}(6.31e^{-5})}$	**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 ≈ 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



Generalization on New Model:

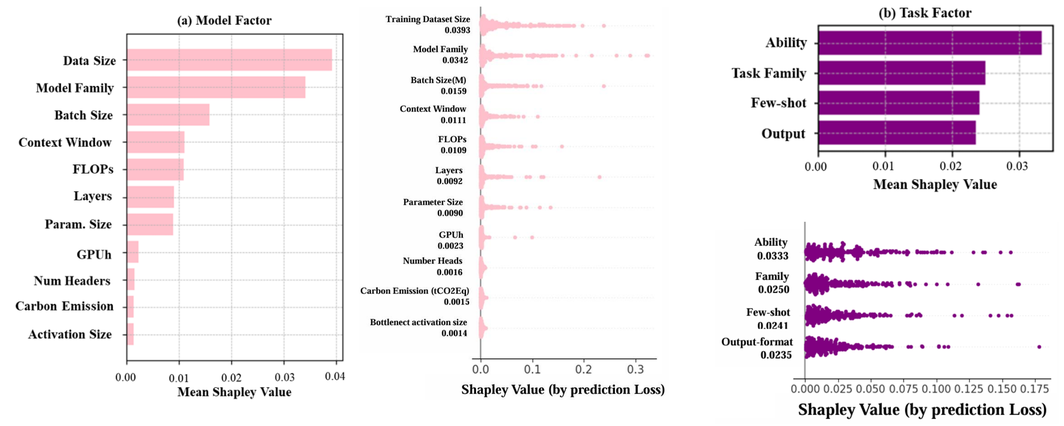
- CPP demonstrates greater adaptability than SL
- Conducting Evaluation on a few tasks can improve Predictability

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



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

Paper Code