ModelLens: Finding the Best for Your Task from Myriads of Models

Rui Cai[1], Wenjie Jacky Mo[1], Xiaofei Wen[1], Qiyao Ma[1], Wenhui Zhu[2], Xiwen Chen[3], Muhao Chen[1], Zhe Zhao[1]

[1]University of California, Davis [2]Arizona State University [3]Morgan Stanley

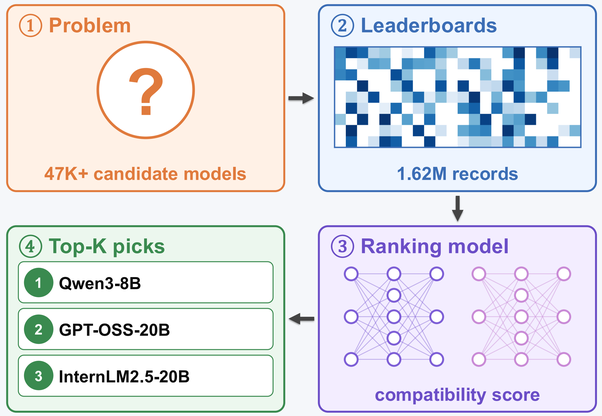
Project page

Code

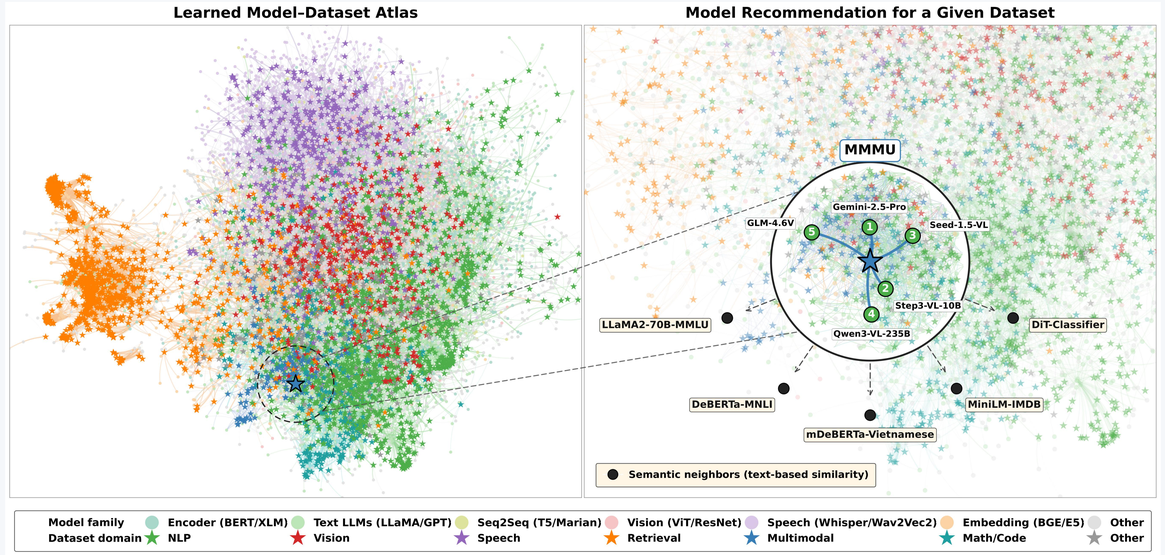
Hundreds of thousands of open models exist, and new models and unbenchmarked datasets appear every day.

Given a new dataset, which model should you use?

ModelLens in four steps



Model recommendation in the wild



Key insight: scattered leaderboard records jointly trace an **atlas of model capabilities**. In this learned map, MMMU's neighbours are competitive multimodal LLMs; text similarity alone finds topically related but unsuitable models.

Scale
47K+ candidates: too many to run on your data

Cold start
new models and new datasets have no records

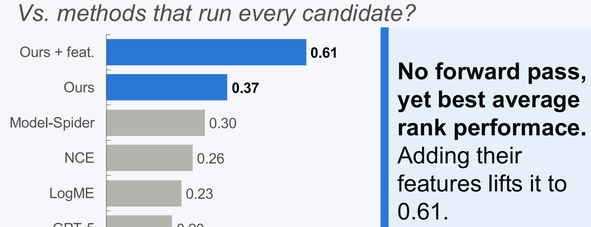
Heterogeneity
tasks and metrics rank models differently

A dataset at scale

1,623,284
evaluation records for training & evaluation

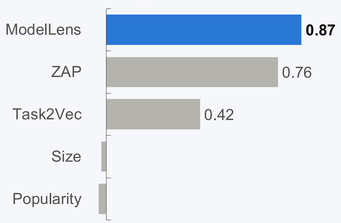
47,062	9,682	2,551	348
models	datasets	tasks	families

vs. standard model-selection benchmark (log scale)



1 Matrix completion

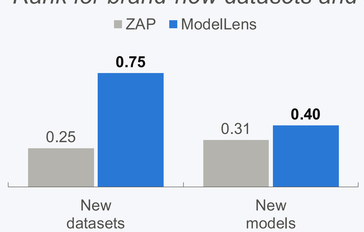
Fill in missing scores on known datasets?



Recovers rankings on 2,967 datasets.
Model size and HF popularity are no better than random ranking.

2 Cold-start generalization

Rank for brand-new datasets and models?



~3× the best baseline on new datasets.
0.75 on 609 held-out datasets; 0.40 on 375 newly released models.

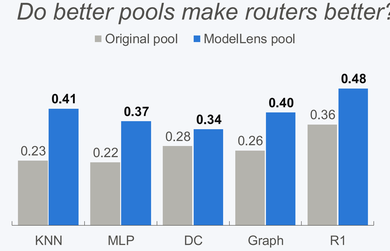
3 Transferability estimation

Vs. methods that run every candidate?

No forward pass, yet best average rank performace.
Adding their features lifts it to 0.61.

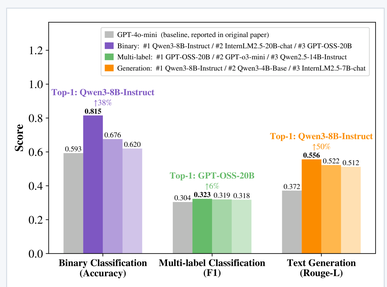
4 Better pools for routing

Do better pools make routers better?

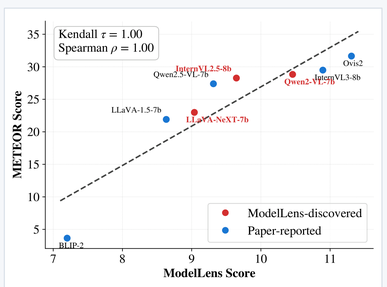


+22% to +81% avg. Exact Match for all 5 routers on 5 QA sets.

5 Case studies: benchmarks released after training



NGQA
text; nutrition-QA
The best model differs per task; ModelLens' top-1 beats the GPT-4o-mini default on all three.



RSVLM-QA
vision-language
Exact ranking of 8 VLMs, incl. 3 models the benchmark never tested (red).