INNOVATE. COLLABORATE. DELIVER.

DATAWorks 2024

A Statistical Framework for Benchmarking Foundation Models with Uncertainty

Giri Gopalan, Los Alamos National Laboratory

April 17th, 2024

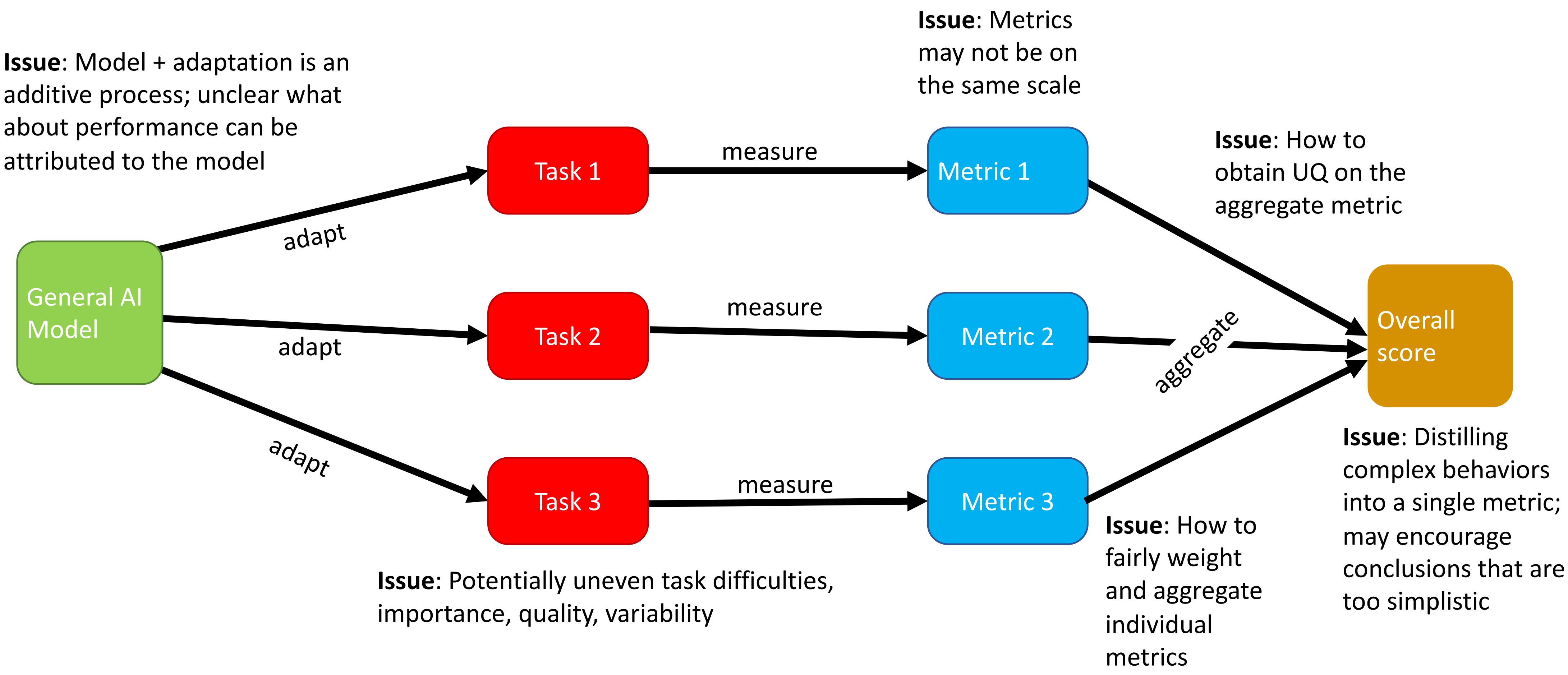
LA-UR-24-23347

Team

- Rachel Longjohn, PhD student at UC Irvine in Statistics and LANL graduate research student.
- Emily Casleton, Scientist in and Deputy Group Leader of the statistical sciences group at Los Alamos.

Workflow Issues

INNOVATE. COLLABORATE. DELIVER.

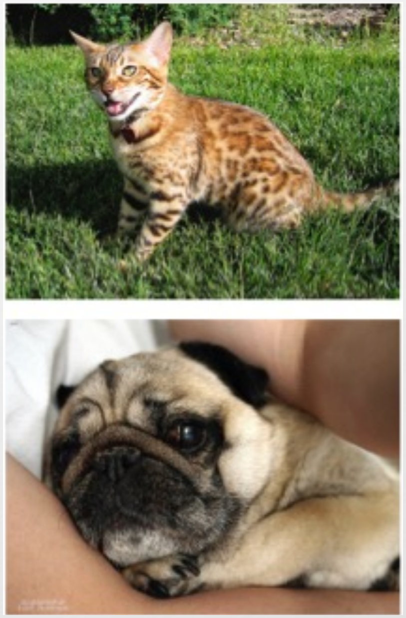


Case Studies

Benchmark	**Modality**	**Metrics**	**Aggregation Mechanism**
VTAB	images	accuracy	unweighted avg.
FLEX	text	accuracy	unweighted avg.
MMLU	text	accuracy	avg. weighted by task size
SuperGLUE	text	accuracy, F1, exact match	unweighted avg.
Xtreme	text	accuracy, F1, exact match	unweighted avg.
BIG-bench	text	accuracy, ECE, Brier score,…	unweighted avg.

VTAB Benchmark

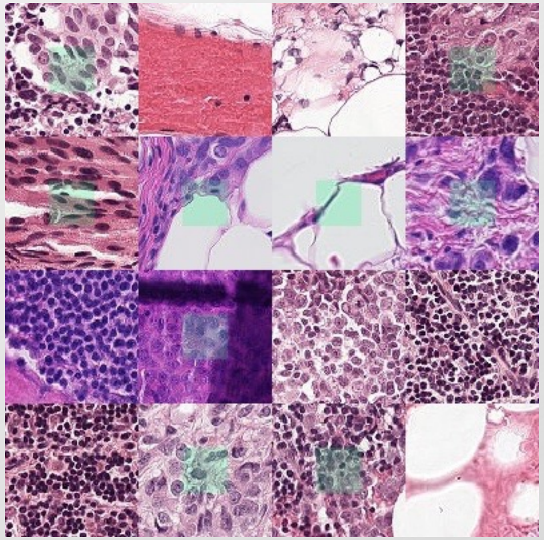
Natural



Parkhi et al., 2012

75564 task items

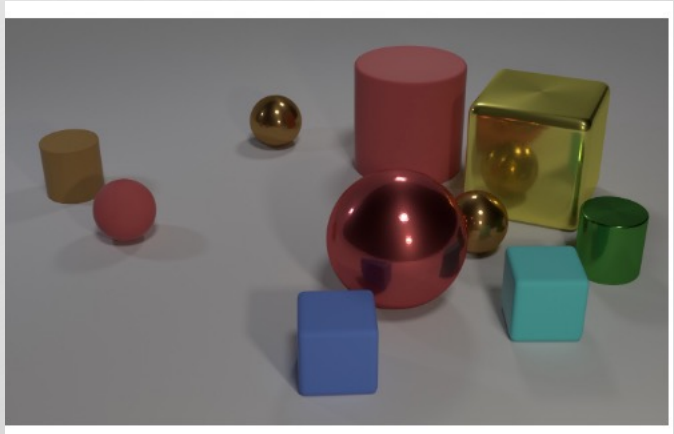
Specialized



Veeling et al., 2018

87138 task items

Structured



Johnson et al., 2017

225202 task items

MMLU Benchmark

- 57 multiple choice tasks across a variety of topics, such as elementary mathematics, U.S. history, computer science, and law
- Grouped into 4 categories

Humanities	Social Sciences	STEM	Other
4705 questions	3077 questions	3153 questions	3107 questions

Existing Leaderboards

VTAB

Model	Natural	Special.	Struct.	Overall
Sup-Rotation-100%	73.6	83.1	55.5	68.0
Sup-Exemplar-100%	73.7	83.1	54.7	67.7
Sup-100%	73.5	82.5	52.1	66.4
Semi-Exemplar-10%	70.2	81.8	52.7	65.3
Semi-Rotation-10%	69.6	82.4	52.5	65.1
Rotation	53.7	78.6	57.3	60.4

MMLU

Model	Human.	Soc. Sci.	STEM	Other	Overall
Codex + REPLUG LSR	76.5	79.9	58.9	73.2	72.6
Codex + REPLUG	76.0	79.7	58.8	72.1	72.1
PaLM 540B	77.0	81.0	55.6	69.6	71.4
Codex	74.2	76.9	57.8	70.1	70.2
Chinchilla	73.1	78.8	55.0	70.3	69.7
LLaMA 65B	61.8	72.9	51.7	67.4	63.2

Adding Bootstrapped 95% Confidence Intervals

	Task 1	Task 2	Task 3	
	Metric 1	Metric 2	Metric 3	Overall score
Bootstrap 1	Metric 1_1	Metric 2_1	Metric 3_1	Agg(Metric i_1)
Bootstrap 2	Metric 1_2	Metric 2_2	Metric 3_2	Agg(Metric i_2)
Bootstrap 3	Metric 1_3	Metric 2_3	Metric 3_3	Agg(Metric i_3)

Adding Bootstrapped 95% Confidence Intervals

- Accuracy
- Test Set Size

	Model 1	**Model 2**	**Model 3**
Q1	correct	correct	incorrect
Q2	correct	incorrect	correct
Q3	correct	correct	correct
…	…	…	…
Q100	correct	correct	incorrect
Acc	95/100	88/100	85/100

VTAB Leaderboard with 95% CIs

Model	Natural	Specialized	Structured	Overall
Sup-Rotation-100%	**73.5** (73.1, 73.9)	**83.2** (82.9, 83.5)	**55.5** (55.1, 55.8)	**68.0** (67.7, 68.2)
Sup-Exemplar-100%	**73.7** (73.3, 74.1)	**83.1** (82.8, 83.3)	**54.7** (54.3, 55.1)	**67.7** (67.5, 68.0)
Sup-100%	**73.4** (73.0, 73.7)	**82.5** (82.2, 82.8)	**52.1** (51.6, 52.6)	**66.3** (66.1, 66.5)
Semi-Exemplar-10%	**70.2** (69.9, 70.6)	**81.8** (81.5, 82.2)	**52.7** (52.3, 53.1)	**65.3** (65.0, 65.6)
Semi-Rotation-10%	**69.5** (69.1, 70.0)	**82.4** (82.1, 82.6)	**52.5** (52.1, 53.1)	**65.1** (64.8, 65.4)
Rotation	**53.7** (53.3, 54.1)	**78.6** (78.3, 78.9)	**57.3** (57.0, 57.8)	**60.5** (60.2, 60.8)

MMLU Leaderboard with 95% CIs

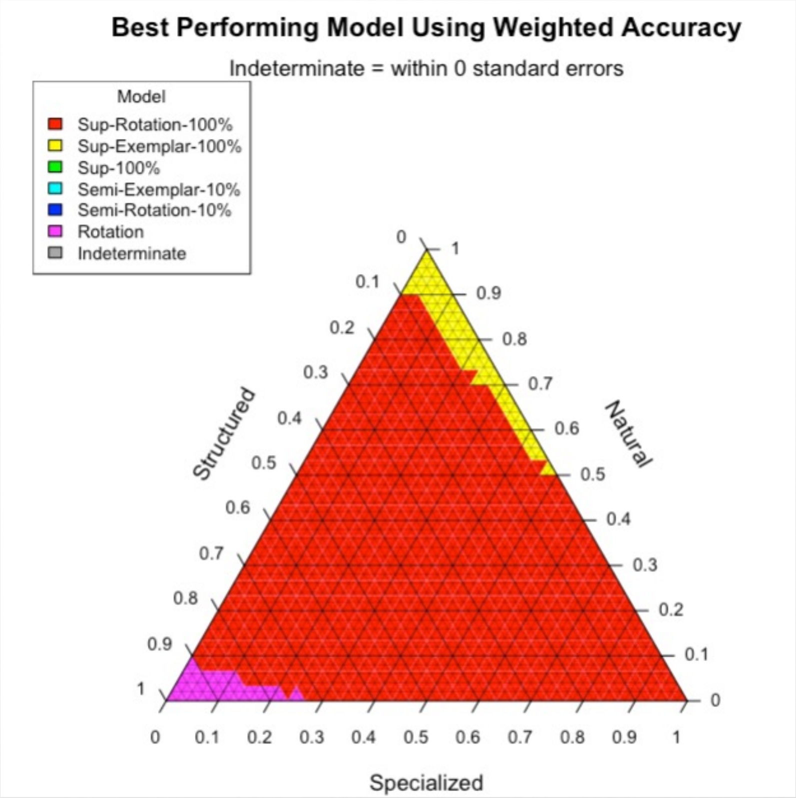
Model	Humanities	Social Sciences	STEM	Other	Overall
Codex + REPLUG LSR	**76.5** (75.5, 77.4)	**79.9** (78.5, 81.4)	**58.9** (57.3, 60.7)	**73.2** (71.4, 74.6)	**72.6** (71.9, 73.2)
Codex + REPLUG	**75.9** (74.8, 77.1)	**79.7** (78.4, 80.8)	**58.7** (56.8, 60.3)	**72.2** (70.8, 73.6)	**72.0** (71.5, 72.7)
PaLM 540B	**77.0** (76.0, 78.1)	**80.9** (79.4, 82.4)	**55.7** (54.3, 57.2)	**69.4** (68.1, 70.9)	**71.4** (70.7, 72.1)
Codex	**74.1** (73.0, 75.3)	**77.0** (75.3, 78.3)	**57.8** (55.7, 59.7)	**70.2** (68.6, 71.8)	**70.2** (69.5, 70.7)
Chinchilla	**73.0** (71.7, 74.4)	**78.7** (77.2, 80.3)	**55.0** (53.3, 56.6)	**70.4** (69.1, 72.0)	**69.6** (69.0, 70.4)
LLaMA 65B	**61.8** (60.3, 63.1)	**73.0** (71.5, 74.5)	**51.7** (50.2, 53.0)	**67.3** (65.7, 68.9)	**63.2** (62.5, 63.9)

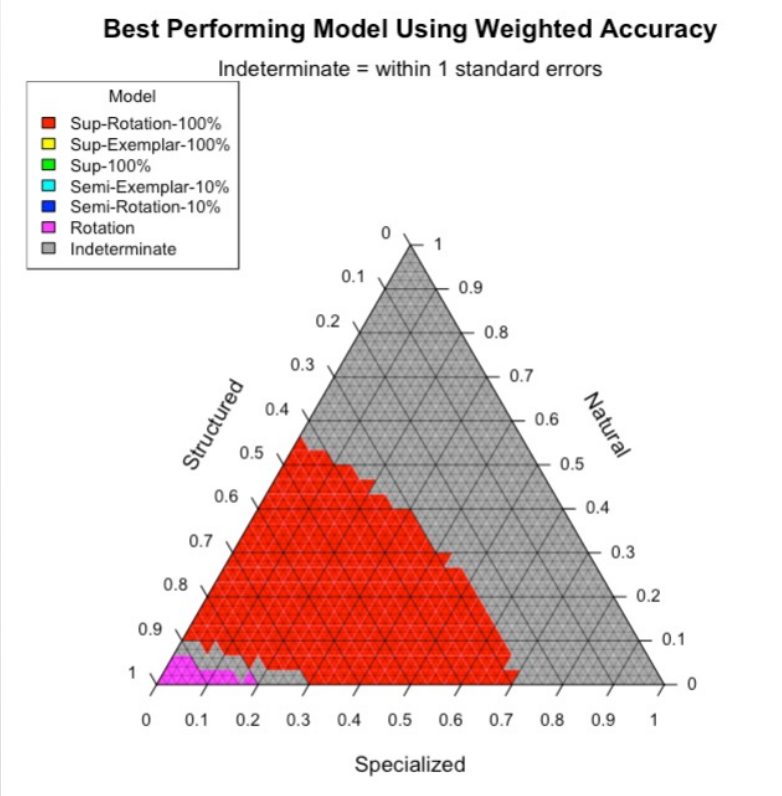
Task Weighting

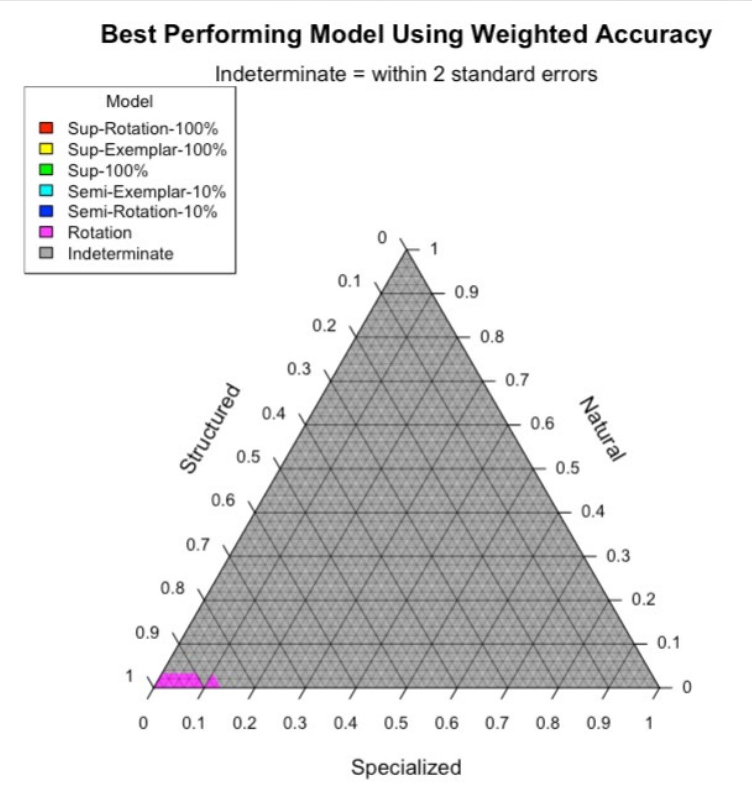
- Many possible reasons why tasks should not be equally weighted when aggregating scores.
- Tasks may have varying levels of importance, quality, difficulty, size, etc.
- What if we want to explore model performance under different weighting mechanisms?

VTAB Performance with Task Weighting

INNOVATE. COLLABORATE. DELIVER.

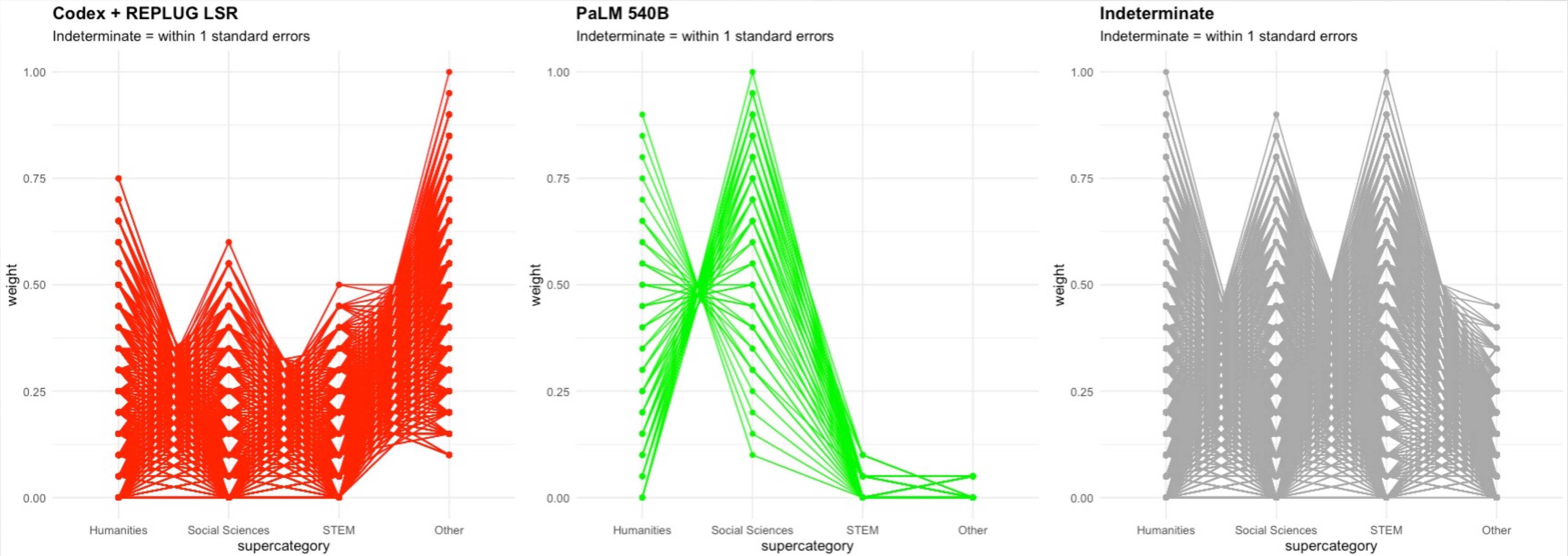






MMLU Performance

INNOVATE. COLLABORATE. DELIVER.



Lawrence Livermore National Laboratory
Los Alamos NATIONAL LABORATORY
OAK RIDGE National Laboratory
Pacific Northwest NATIONAL LABORATORY
Sandia National Laboratories

Bayesian Hierarchical Modeling for Evaluation Data

- An alternative to the bootstrap for quantifying uncertainty is to use a Bayesian hierarchical model (BHM).
- Mentioned for classifiers in *Active Bayesian Assessment of Black-Box Classifiers* by Ji, Logan, Smyth, and Stevyers, but have not seen an implementation.
- Allows for borrowing strength across distinct tasks for assessing a given foundation model. E.g., common underlying distribution for foundation model accuracy across distinct tasks.
- Can quantify uncertainty over FM performances by sampling from the posterior predictive distribution of task performances.

BHM for MMLU and VTAB

- $\boldsymbol{Y}_{ij}$ is the number of correct responses for foundation model ***i*** on task ***j***.
- $\boldsymbol{\theta}_{ij}$ is the probability of accurate response for foundation model ***i*** on task ***j***.

$$Y_{ij}|\theta_{ij} \sim Binom(\theta_{ij}, N_j)$$

is the data model, and the model for θ_{ij} is

$$\theta_{ij}|\alpha_i, \beta_i \sim Beta(\alpha_i, \beta_i).$$

The BHM specification is completed by putting a distribution on α_i, β_i:

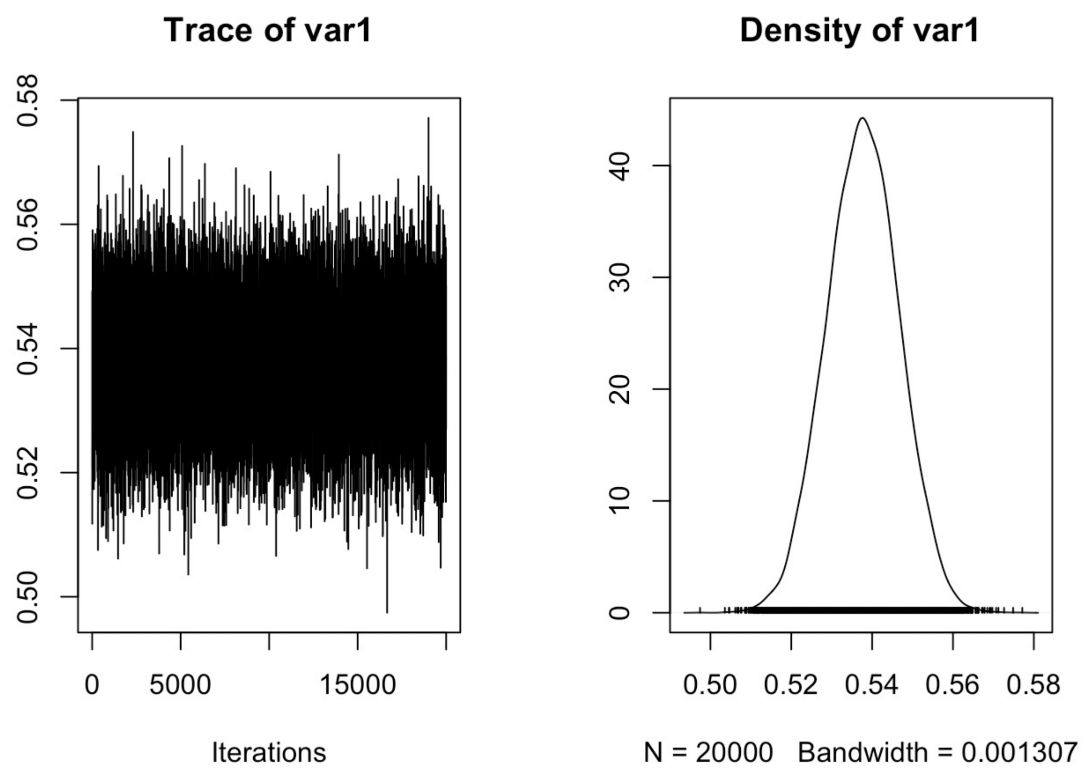
$$log(\alpha_i) \sim Normal(\mu_\alpha, \sigma_\alpha)$$

and

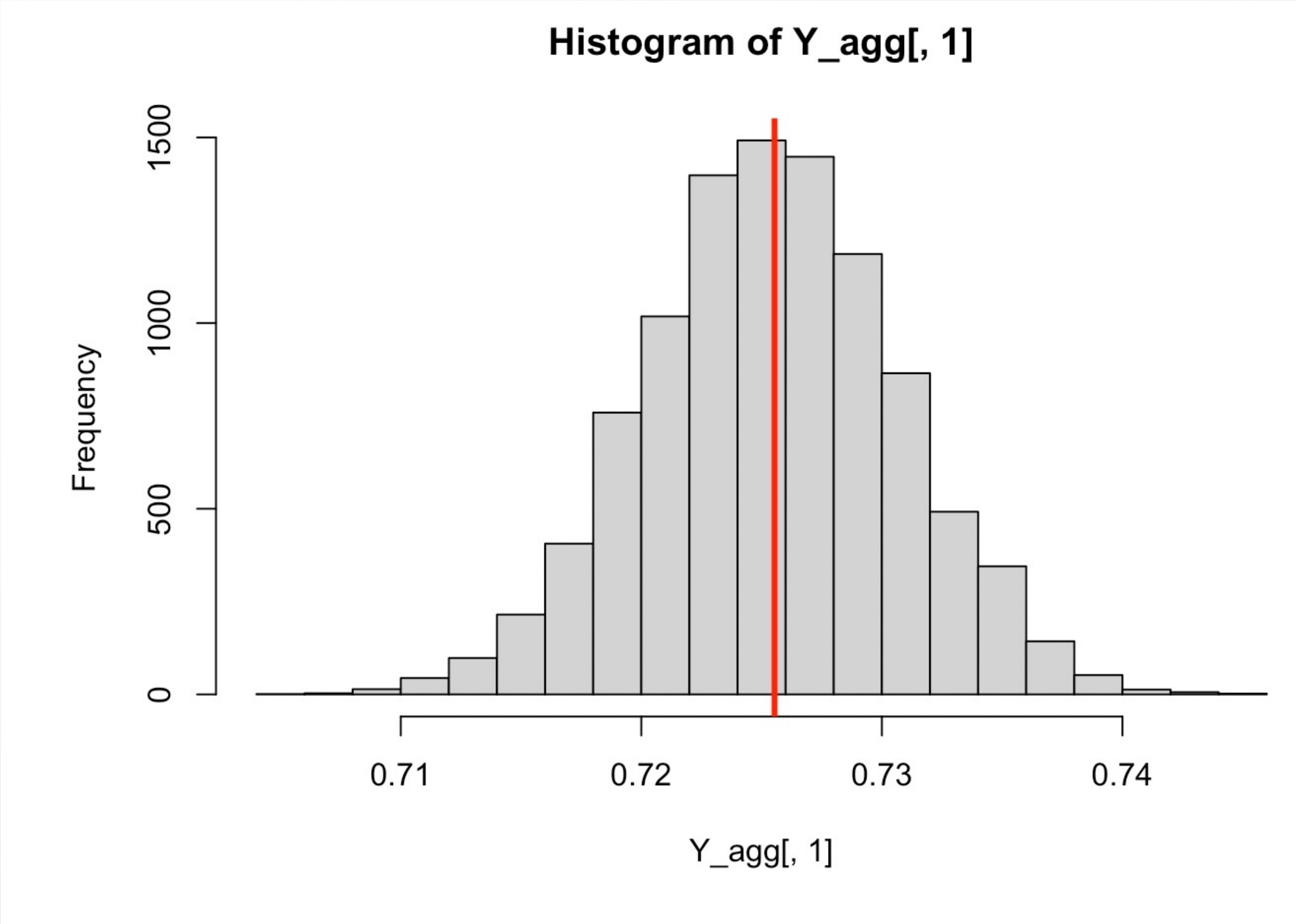
$$log(\beta_i) \sim Normal(\mu_\beta, \sigma_\beta)$$

Fit the BHM with Markov Chain Monte Carlo


```
#pick a random theta_ij
plot(as.mcmc(theta_samps[11,2,]))
```



Posterior Predictive Check



MMLU: BHM vs. Bootstrap

INNOVATE. COLLABORATE. DELIVER.

Model	BHM	Bootstrap
Codex + REPLUG LSR	(71.5, 73.5)	(71.9, 73.2)
Codex + REPLUG	(71.0, 73.1)	(71.5, 72.7)
PaLM 540B	(70.4, 72.4)	(70.7, 72.1)
Codex	(69.1, 71.2)	(69.5, 70.7)
Chinchilla	(68.6, 70.7)	(69.0, 70.4)
LLaMA 65B	(62.1, 64.3)	(62.5, 63.9)

VTAB: BHM vs. Bootstrap

Model	BHM	Bootstrap
Sup-Rotation-100%	(67.1, 68.9)	(67.7, 68.2)
Sup-Exemplar-100%	(66.7, 68.6)	(67.5, 68.0)
Sup-100%	(65.4, 67.2)	(66.1, 66.5)
Semi-Exemplar-10%	(64.3, 66.2)	(65.0, 65.6)
Semi-Rotation-10%	(64.1, 65.9)	(64.8, 65.4)
Rotation	(59.5, 61.4)	(60.2, 60.8)

UQ: BHM vs. Bootstrap

- BHM produces slightly wider intervals than bootstrap, but generally agree.

- Empirically, we see very close agreement with more data in both VTAB and MMLU.

Standardizing Scores

Standardized score = (raw - low)/(high-low)

- Standardizing Task Evaluations
  - Primary advantage: allows one to transform general evaluation scores to [0,1] and apply the same UQ framework presented for accuracies.
  - Can bootstrap, BHM, and visualize in the same manner after standardizing!
  - Caution: could still make sense to re-weight even after standardizing.

Rank Aggregation Across Tasks

1. Borda count (rank by average rank).
2. Kemeny consensus.
3. Bayesian posterior rank probabilities.

Rank by Average Rank (Borda Count)

- For each task, rank the foundation models, average ranks across tasks, and then rank models by average rank. (UQ via bootstrap or BHM.)

MMLU

Model	95% Conf. Interval
Codex + REPLUG LSR	(1.25 – 2.50)
Codex + REPLUG	(1.50 – 3.00)
PaLM 540B	(2.25 – 3.50)
Codex	(3.25 – 4.50)
Chinchilla	(3.63 – 4.75)
LLaMA 65B	(6.00 – 6.50)

VTAB

Model	95% Conf. Interval
Sup-Rotation-100%	(2.00 – 3.33)
Sup-Exemplar-100%	(2.00 – 2.33)
Sup-100%	(3.33 – 4.67)
Semi-Exemplar-10%	(4.33 – 5.33)
Semi-Rotation-10%	(3.00 – 4.33)
Rotation	(4.33 – 4.67)

Kemeny Consensus

- Given a list of ranks, find a consensus rank that minimizes the average Kendall distance (i.e., number of inversions) between the consensus and the list of ranks.
- Kendall distance examples:
  - D((1,2,3,4), (4,1,2,3)) = 3
  - D((3,1,2), (2,1,3)) = 1
  - D((5,3,2,4,1), (4,2,1,5,3)) = 3
- While Kemeny consensus satisfies some nice properties, **it is not unique**, which is a source of uncertainty beyond sampling variability.

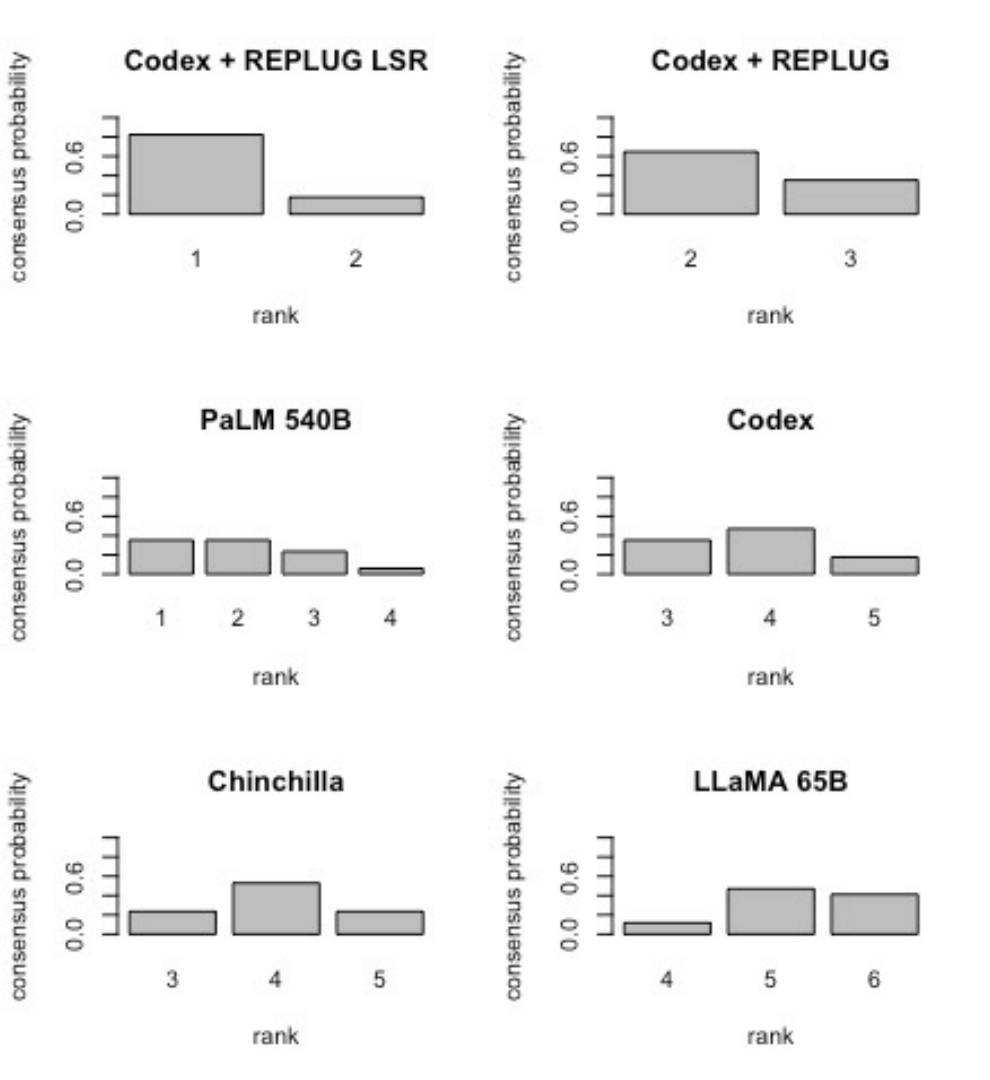
MMLU Kemeny Consensus

$Consensus

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]
[1,]	1	2	3	4	5	6	7	8	9	10	11	12
[2,]	1	2	3	4	4	5	6	7	8	9	10	11
[3,]	1	2	3	5	4	6	7	8	9	10	11	12
[4,]	1	2	3	3	4	5	6	7	8	9	10	11
[5,]	1	2	4	3	5	6	7	8	9	10	11	12
[6,]	1	2	2	3	4	5	6	7	8	9	10	11
[7,]	1	2	2	3	3	4	5	6	7	8	9	10
[8,]	1	2	2	4	3	5	6	7	8	9	10	11
[9,]	1	3	2	4	5	6	7	8	9	10	11	12
[10,]	1	3	2	4	4	5	6	7	8	9	10	11
[11,]	1	3	2	5	4	6	7	8	9	10	11	12
[12,]	1	2	1	3	4	5	6	7	8	9	10	11
[13,]	1	2	1	3	3	4	5	6	7	8	9	10
[14,]	1	2	1	4	3	5	6	7	8	9	10	11
[15,]	2	3	1	4	5	6	7	8	9	10	11	12
[16,]	2	3	1	4	4	5	6	7	8	9	10	11
[17,]	2	3	1	5	4	6	7	8	9	10	11	12

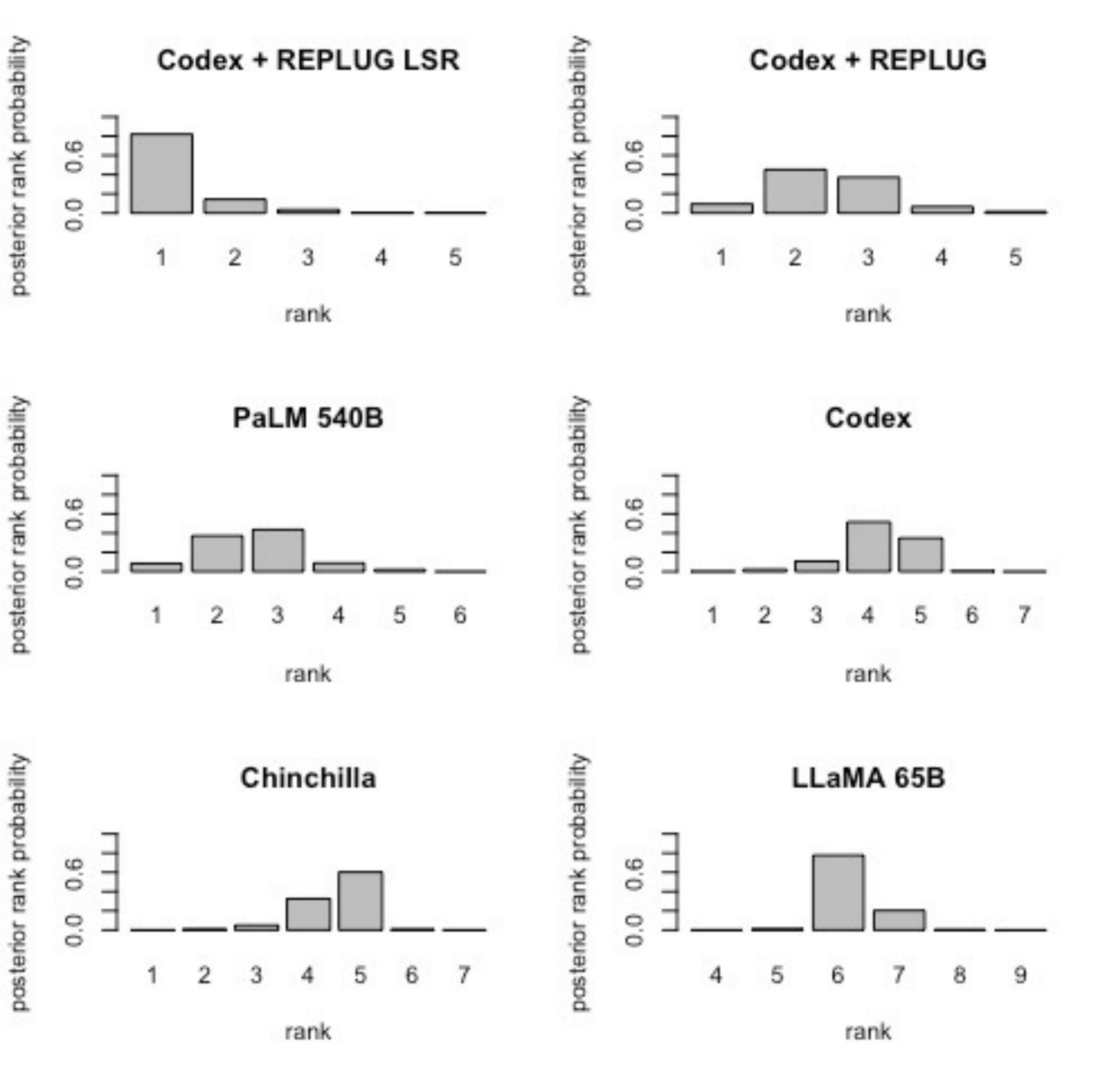
Via **ConsRank R** package

MMLU Kemeny Consensus - UQ



Barplot illustrates proportion of 17 consensus rankings for which the model is ranked i (i on x axis), i.e., a *consensus probability.*

MMLU Bayesian Ranking Alternatives



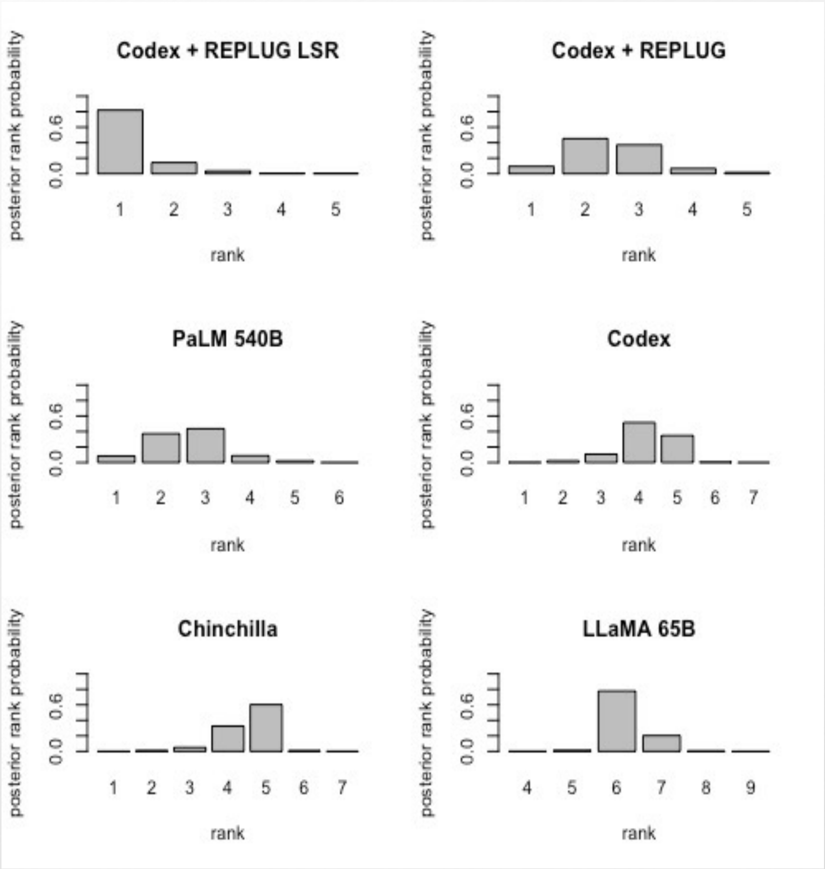
Fit a Bayesian *Thurstone-Mosteller-Daniels* latent variable model via **BayesRankAnalysis R** package.

Plot posterior rank probabilities derived from MCMC output. Some similarities with consensus probabilities but not quite the same, and more uncertainty.

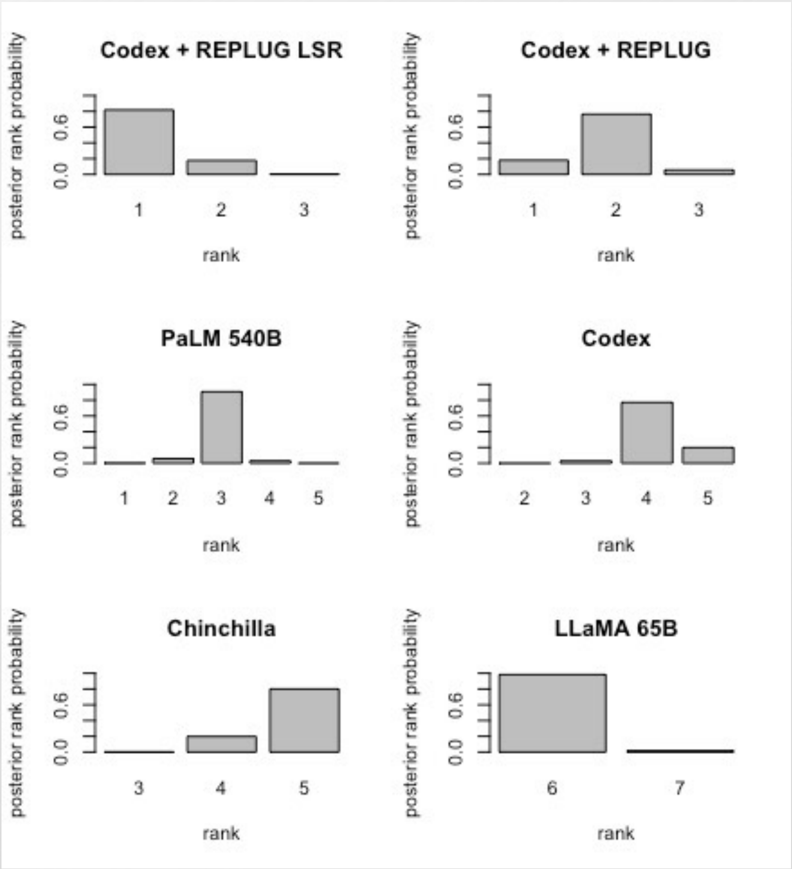
NOTE: Can extract similar posterior probability plots via the BHM, ranking based on accuracy probabilities.

BHM Posterior Over Ranks

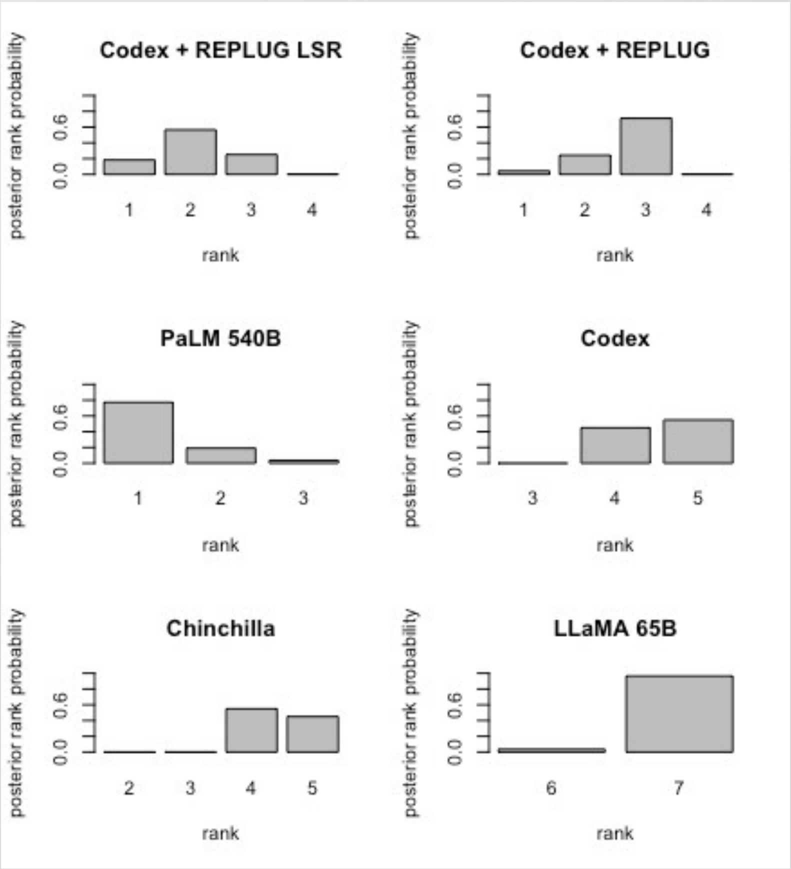
Rank based on the estimand $\sum_j w_j \theta_{ij}$



Thurstone-Mosteller-Daniels



equal weights



hum weight = .5, social science weight = .45, STEM weight = .025, other weight = .025.

Future Work

- Predictive Uncertainty
  - Target only a small portion of the network for tractability (e.g., fine tuning).
  - Split conformal inference – does not require iterative refitting.
  - Resampling techniques.
  - Random ensemble methods.

Thank you