Stronger Random Baselines for In-Context Learning

Gregory Yauney
Cornell University & USC

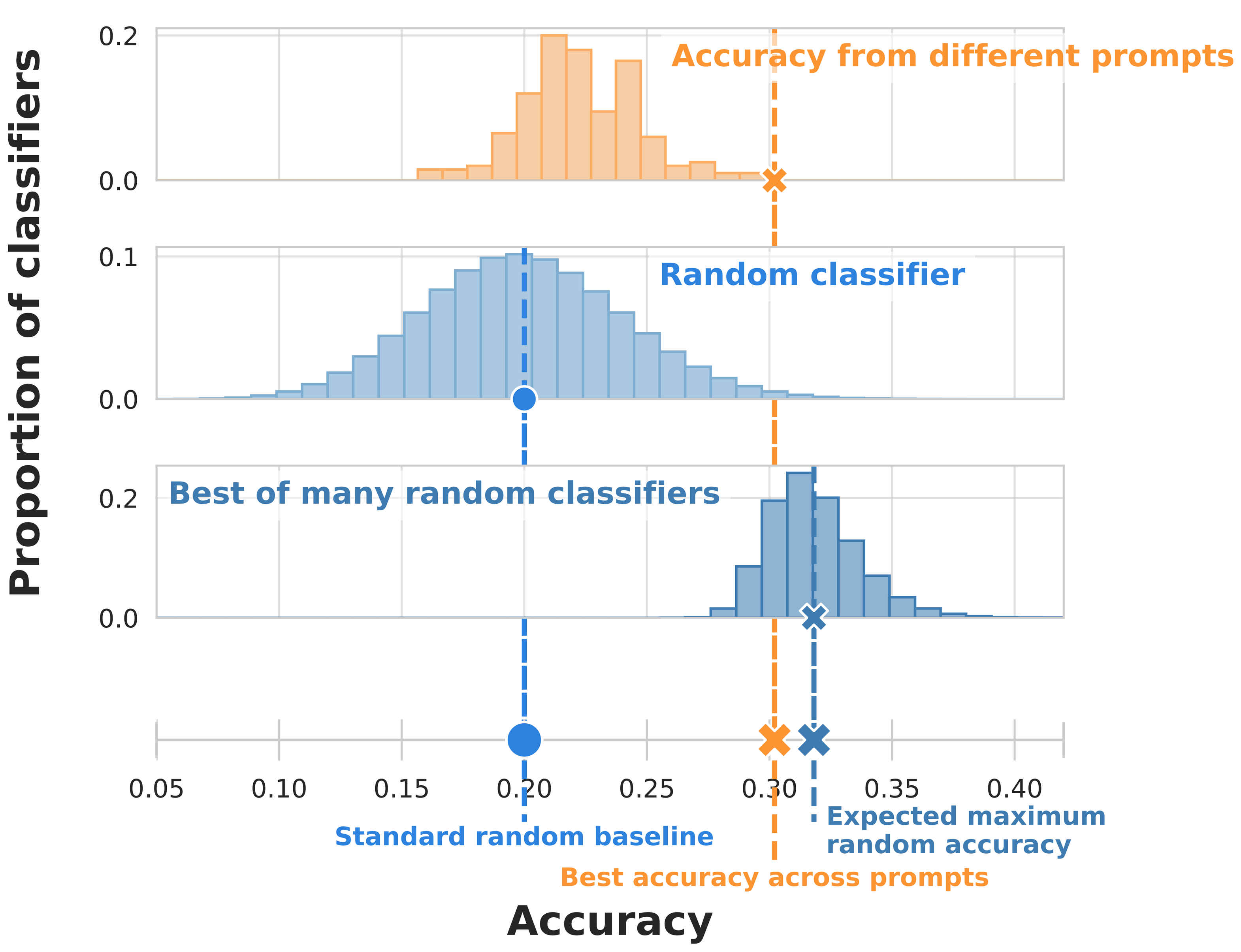
David Mimno
Cornell University

Random baselines can overstate a model's classification performance when validation sets are small and reused.

Goal: reduce premature test set usage by better contextualizing validation accuracy

Solution: **a drop-in replacement baseline** that compares to the expected best accuracy from among *many* different random classifiers

Example: pick the best prompt from among 200 on a dataset of n = 100 examples with 5 label choices



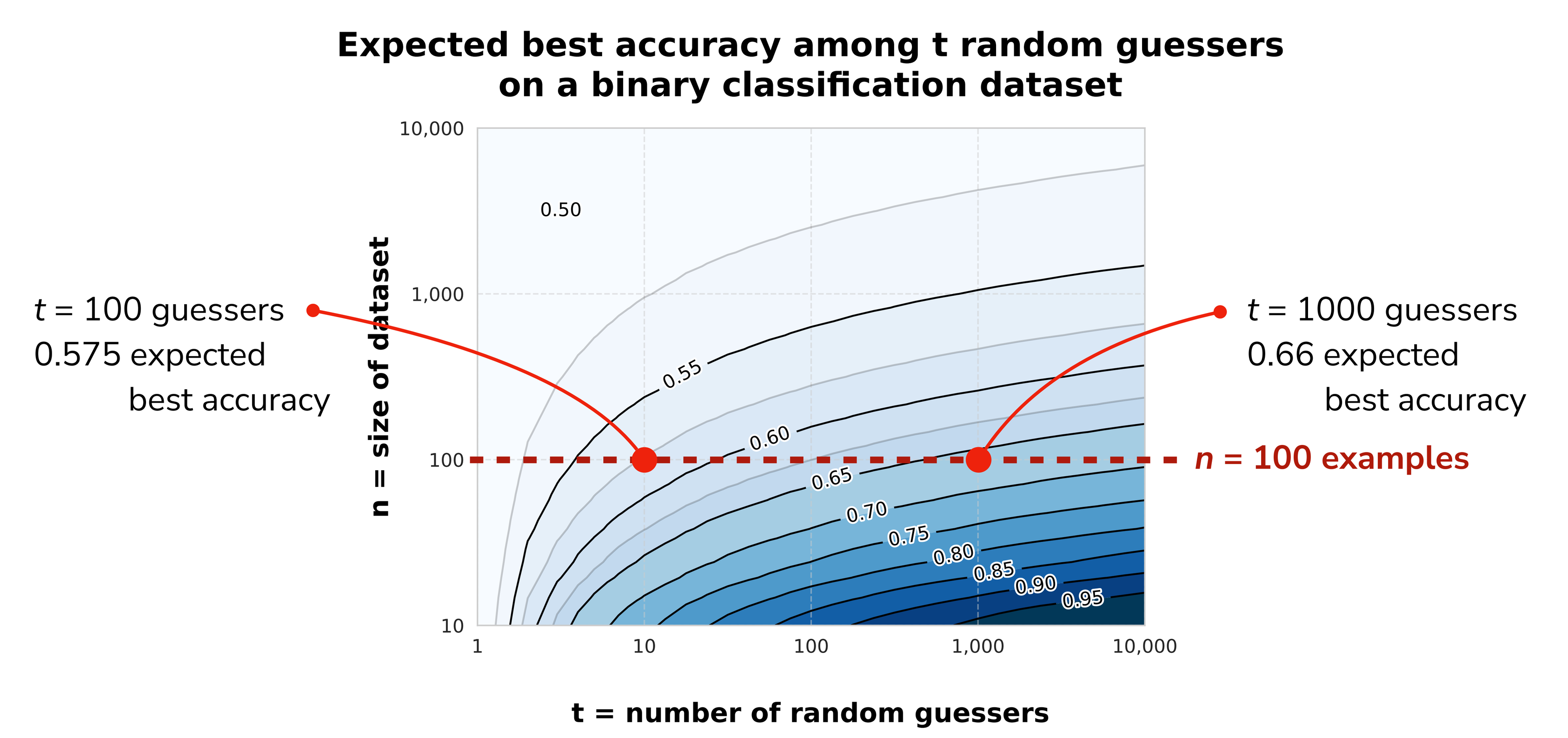
The best prompt outperforms one random guesser but not the best of 200 random guessers in this example.

ICL evaluations:
1. many **small** datasets
2. **challenging** tasks
3. extensive validation **dataset reuse** due to prompt variability

Standard random baseline: expected accuracy of one random classifier

Maximum random baseline: expected *best* accuracy among *t* random classifiers that guess uniformly at random, independently across examples

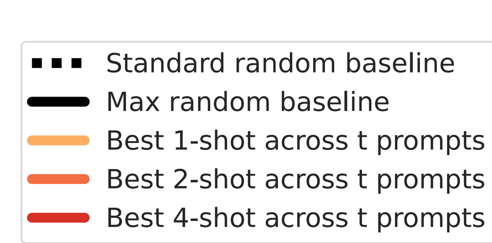
calculate using the maximum order statistic of a binomial distribution

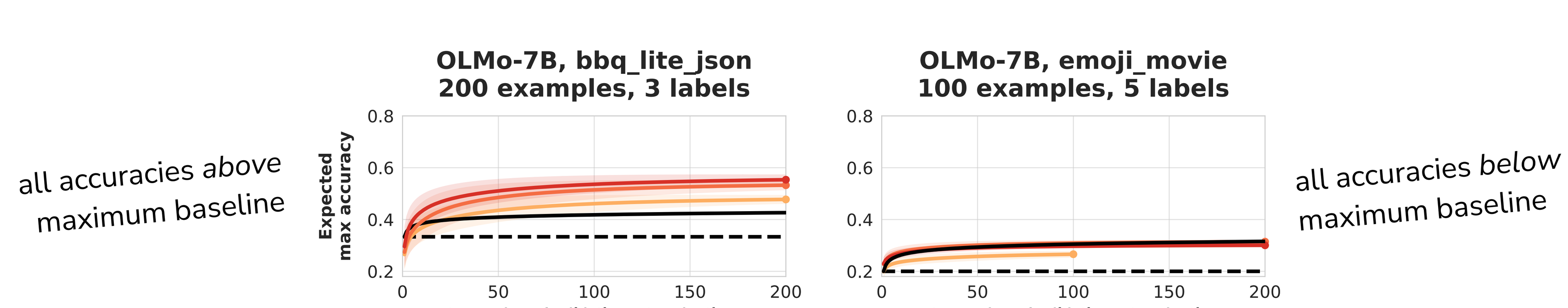


Baselines differ for prompt selection

Setting: pick the best prompt from 200 with different orderings and choices of demonstrations

7B models: Llama-2, OLMo, Falcon, Alpaca, OLMo-Instruct, Falcon-instruct





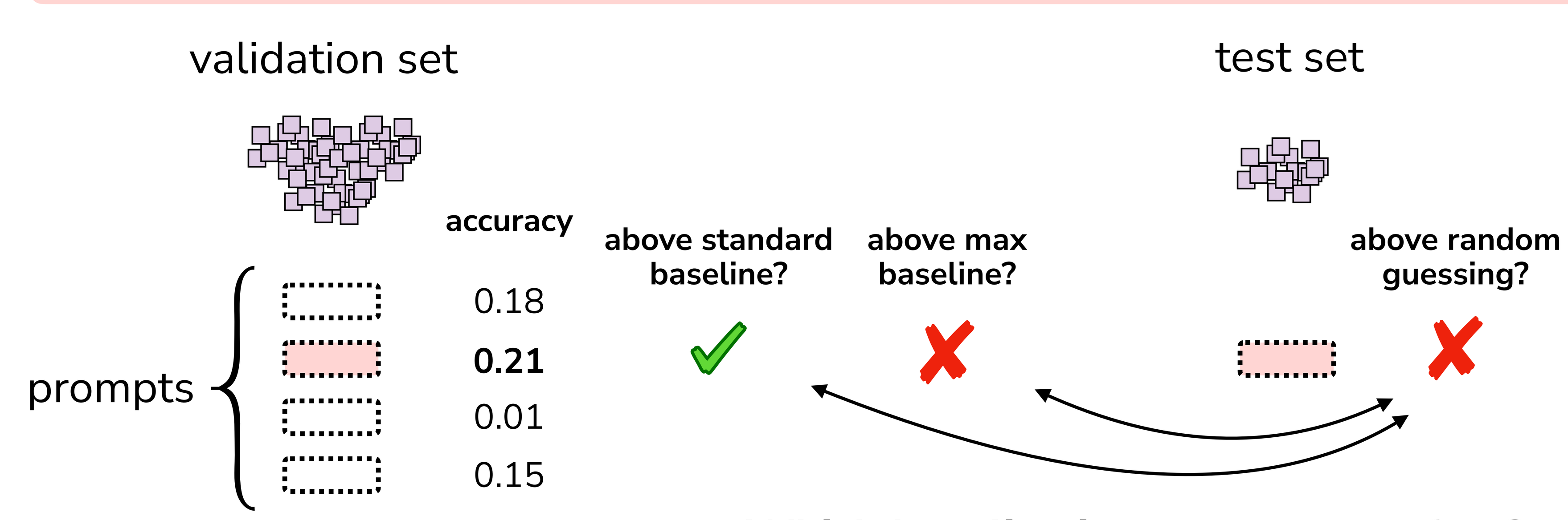
		Base model									Instruction-tuned								
		1-shot			2-shot			4-shot			1-shot			2-shot			4-shot		
BIG-bench Lite dataset	n	L	O	F	L	O	F	L	O	F	A	O	F	A	O	F	A	O	F
novel_concepts	32	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
known_unknowns	46	●	◐	◐	●	●	●	●	●	◐	◐	◐	◐	●	◐	◐	●	●	◐
code_line_description	60	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
emoji_movie	100	●	◐	◐	●	◐	◐	●	◐	●	◐	◐	◐	●	◐	◐	●	●	◐
conceptual_combinations	103	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
strange_stories	174	●	●	●	●	●	●	○	○	○	●	●	●	●	●	●	●	○	○
hindu_knowledge	175	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
bbq_lite_json	200	●	●	●	●	●	●	●	●	●	●	●	○	●	●	●	●	●	●
formal_fallacies_syllogisms_negation	200	◐	◐	◐	◐	◐	◐	○	○	○	◐	◐	◐	●	◐	◐	◐	○	○
language_identification	200	●	◐	◐	●	◐	◐	◐	○	◐	●	◐	●	◐	◐	●	◐	◐	◐
logical_deduction	200	●	●	●	●	●	●	●	●	◐	●	●	●	●	●	●	●	◐	◐
play_dialog_same_or_different	200	●	●	●	○	◐	○	○	○	○	●	●	●	●	●	●	●	○	○
strategyqa	200	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
symbol_interpretation	200	○	◐	◐	○	○	○	○	○	○	◐	◐	◐	◐	○	○	◐	○	○
vitaminc_fact_verification	200	◐	●	●	●	●	●	○	●	○	●	●	●	●	●	●	●	●	○
winowhy	200	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

○ below standard baseline, below maximum baseline
◐ above standard baseline, below maximum baseline
● above standard baseline, above maximum baseline

More than a fifth of results that are above the standard baseline are not above the max baseline:

$$\frac{◐}{◐ + ●} = 22\%$$

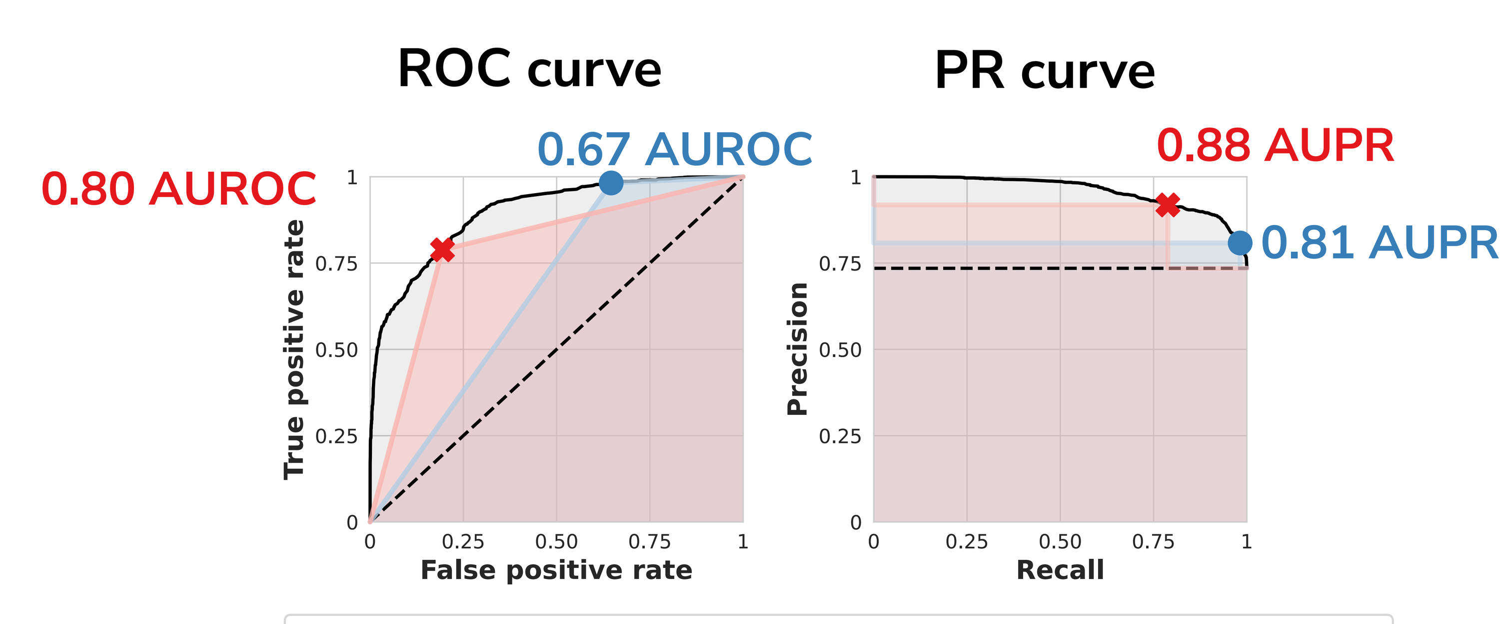
The maximum random baseline generalizes better



Which baseline is correct more often?

repeat for 100 trials per task/model combo
aggregate across tasks + models

Beating the maximum random baseline on validation is a better indicator of test performance across 16 BIG-bench Lite tasks.



Both baselines induce the same ROC and PR curves but each specifies a different point on them.

If your best prompt doesn't outperform the maximum baseline on the validation set, do not evaluate on the test set.

```
pip install max-random-baseline
```