Data Similarity is Not Enough to Explain Language Model Performance

Gregory Yauney
Cornell University
gyauney@cs.cornell.edu

Emily Reif
Google Research
ereif@google.com

David Mimno
Cornell University
mimno@cornell.edu

Why do language models perform better on some tasks than others?

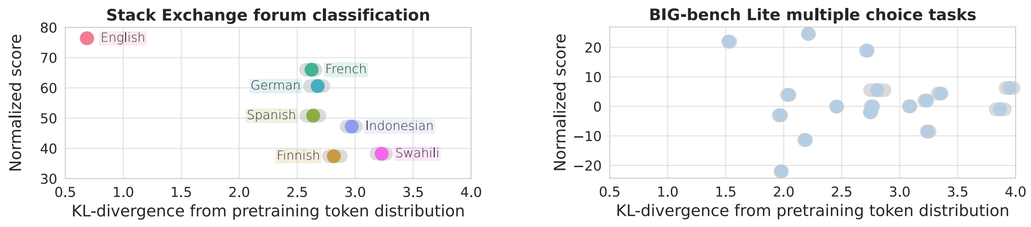
task-level similarity hypothesis — A language model will perform better on a task when the task's data distribution is more similar to the model's pretraining data distribution.

Why do language models perform better on some examples than others?

example-level similarity hypothesis — A language model will perform better on an example when the example is more similar to the examples in the model's pretraining data distribution.

Is there evidence that either of these broad similarity hypotheses account for the performance variation across tasks and examples that we can directly measure?

In some carefully controlled settings, but not in general!



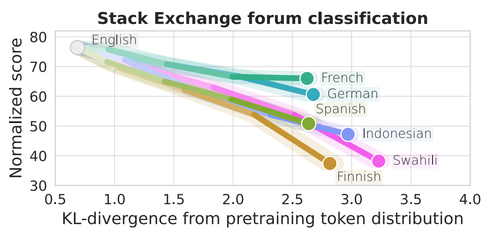
Similarity explains performance in a controlled setting

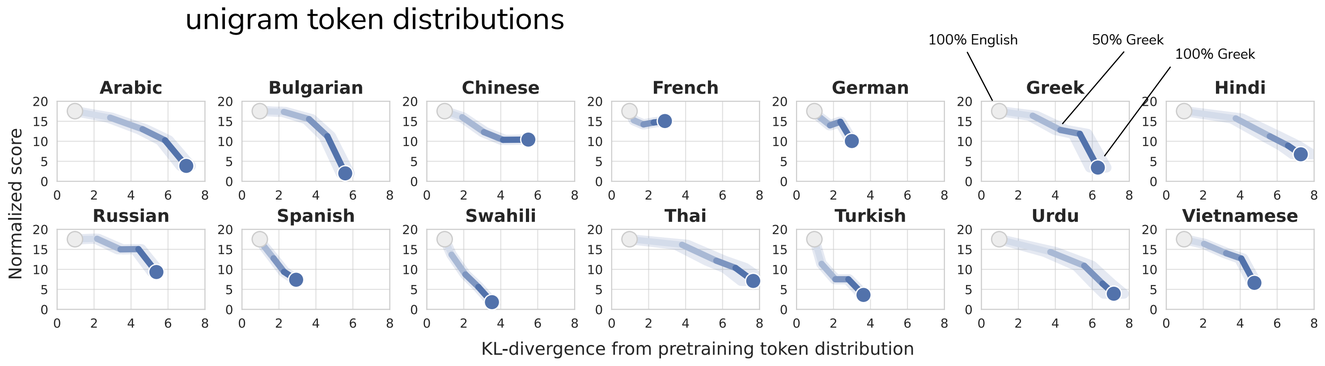
Idea: Fix the task, change the dataset

Tasks: 1. Stack Exchange forum classification
2. XNLI

Data variants: same documents in multiple languages at various stages of translation

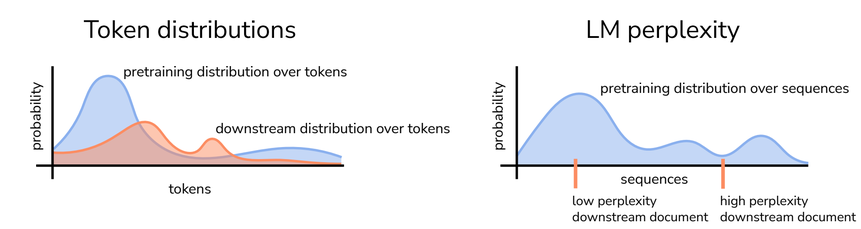
Similarity: KL-divergence between unigram token distributions





Multiple ways of measuring textual similarity

between pretraining dataset and downstream datasets



Dataset similarity does not determine performance across tasks

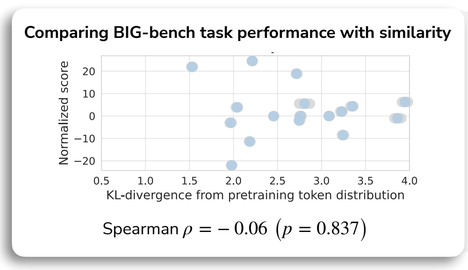
Tasks: BIG-bench Lite multiple choice

Models: Pythia, pretrained on the Pile
T5, pretrained on C4
Flan-T5, pretrained on C4

Setting: zero-shot and few-shot

If the similarity hypothesis is true, we expect to see significant Spearman rank correlations between task similarity to the pretraining data and performance

But we don't! After accounting for multiple comparisons



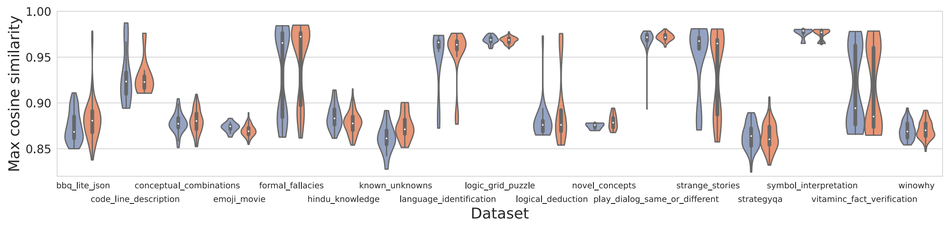
Spearman ρ between performance and similarity to pretraining data

	Pythia-6.9B 0-shot		T5 v1.1 XL 0-shot		T5 v1.1 XL 2-shot		Flan-T5 XL 0-shot		Flan-T5 XL 2-shot	
Bigram KL-divergence (-)	-0.06	0.837	-0.10	0.708	-0.04	0.897	-0.67	0.005	-0.64	0.007
MAUVE score (+)	-0.36	0.165	0.19	0.478	0.27	0.318	0.26	0.329	0.46	0.075
Max cosine similarity (+)	0.08	0.778	-0.10	0.716	-0.26	0.339	0.14	0.594	-0.13	0.633
Mean cosine similarity (+)	0.07	0.795	-0.05	0.867	-0.19	0.478	0.19	0.492	-0.10	0.713
Input perplexity (-)	0.09	0.729	-0.13	0.644	-0.24	0.374	-0.16	0.542	0.12	0.664
Correct target perplexity (-)	0.44	0.085	0.09	0.745	-0.02	0.948	-0.21	0.431	-0.25	0.356

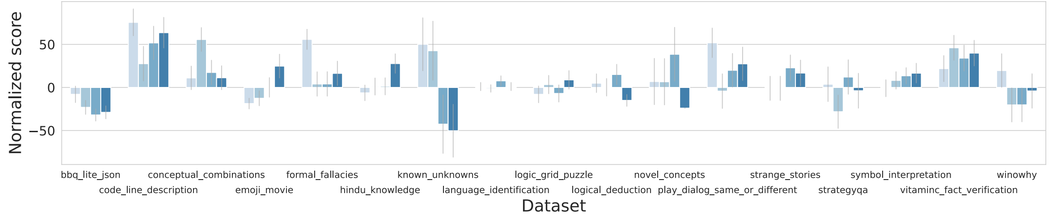
Dataset similarity does not determine performance across examples

Compare each downstream example to the **entire** pretraining dataset: calculate each example's maximum cosine similarity to any pretraining document

Correct examples are not significantly more similar than incorrect examples



More similar examples do not afford higher performance than less similar examples



Which hypotheses are consistent with these results?

1. The similarity measures we tried do not capture the true variation in language that accounts for performance variation.

   We need new measures of textual similarity

2. Existing benchmarks might already be so similar to web-scale pretraining datasets that other factors determine performance.

3. Data similarity is not as important a factor in language model performance as commonly assumed.

   We need to go beyond data similarity to explain task and example difficulty with approaches like training data attribution

How do these results fit in with related work?

- LMs perform better on factoid QA examples with entities frequently found in the pretraining dataset. We ask a **harder question**: will a model perform well on an example if that example is similar to any pretraining document?
- It's clear that **pretraining data matters**! But our results make it increasingly unlikely that similarity broadly construed is the most determinative factor for task performance

Paper: arxiv.org/pdf/2311.09006.pdf

Code + data: github.com/gyauney/data-similarity-is-not-enough