Future Challenges for Text-as-Data in Economics

Stephen Hansen (UCL)

Relevant papers

Text Algorithms in Economics (with E. Ash), 2023, *Annual Review of Economics*

Remote Work across Jobs, Companies, and Space (with P. Lambert, N. Bloom, Y. Muvdi, R. Sadun, B. Taska), 2023, WP

Inference for Regression with Variables Generated from Unstructured Data (with L. Battaglia, T. Christensen, and S. Sacher), 2024, WP

Where are We Now?

Text (and other unstructured data) measures otherwise unobservable variables important for understanding the economy:

- Newspaper text → uncertainty
- Corporate filings → product market competition
- Job postings → skills
- Satellite images → local economic development

Mostly reduced-form analysis, but early examples of input into structural models

How Do We Extract Information From Text?

Text is sequential data $w = (w_1,\ldots,w_N)$.

Bag-of-Words Model: represent data as count vector over words

- Count frequency of specific terms
- Compare distance between count vectors
- Reduce dimensionality of term counts with topic models

But all these ignore sequential structure

Modern Developments

Text is sequential data $w = (w_1,\ldots,w_N)$

Neural Language Models: Build vector representations of words within neural networks targeted at word prediction problems.

- Word embeddings: word2vec, GloVe
- Sequence embeddings: BERT, RoBERTa
- Large Language Models: GPT family, Gemini, Claude, etc.

We now have excellent models for $Pr[w_{N+1} \mid w]$

Important Questions for Field

Crucial issue is that economic researchers facing this landscape have many choices to make with little guidance.

1. **Which model should we use**?

2. **How should we perform inference given a model**?

Which model should we use?

Important to define relevant measurement problem

1. Distance between documents
2. Concept detection
3. Relationship between concepts
4. Relating text to metadata

LLM may or may not be appropriate, e.g. distance vs concept detection

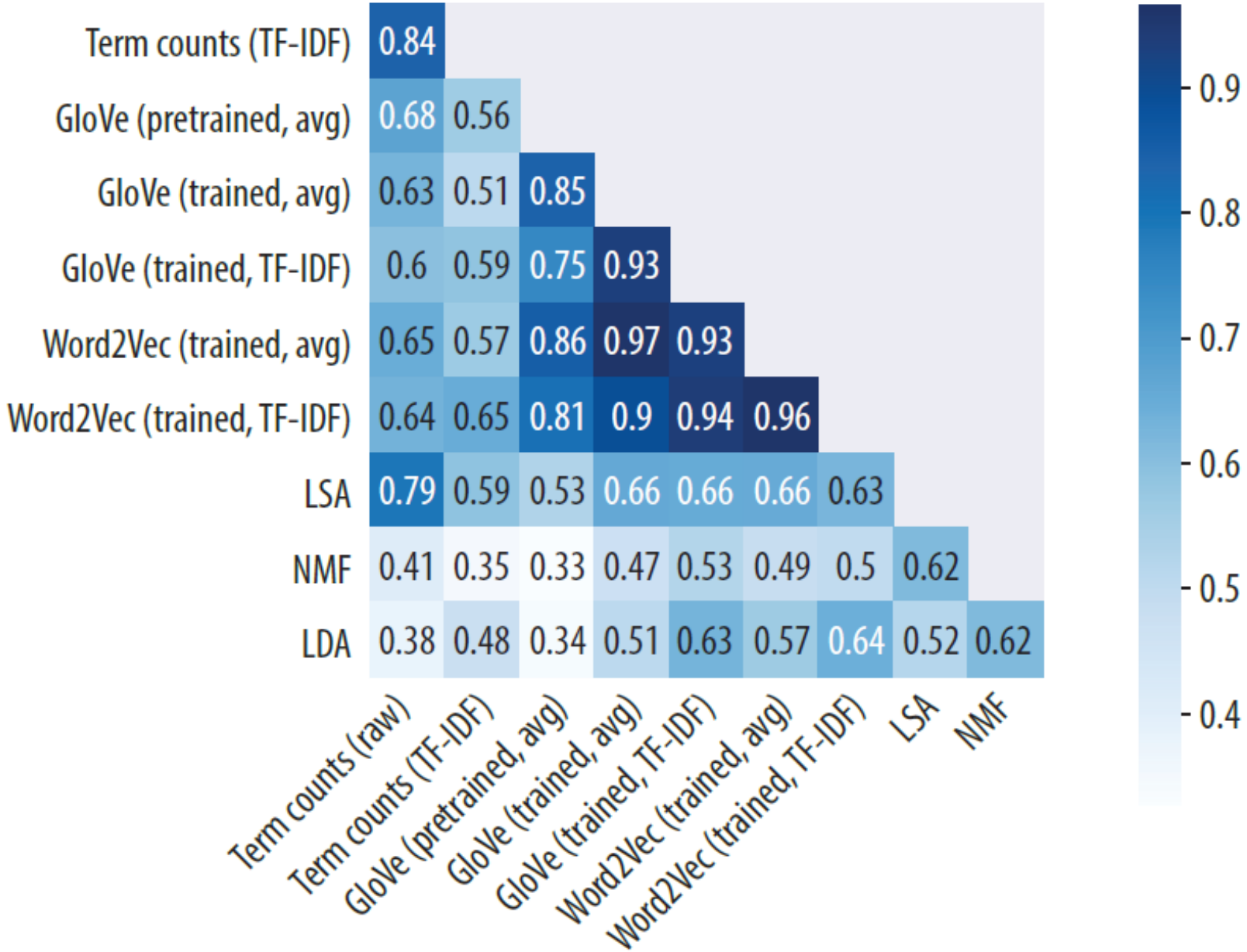
How to trade-off accuracy vs other goals: transparency, replicability, interpretability

Does the Choice of Model Matter?

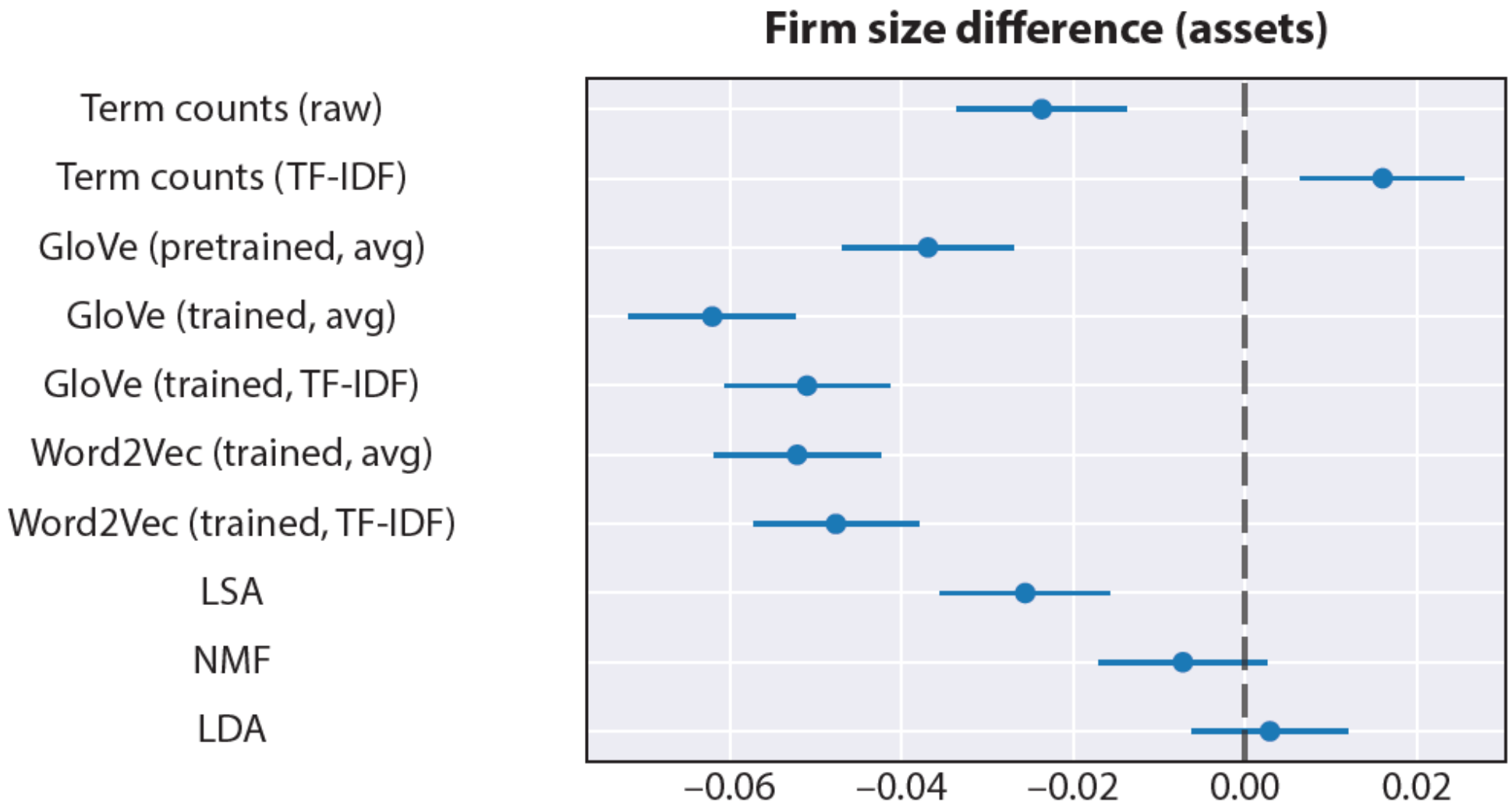
Consider the problem of comparing document similarity in standard corpus: *Risk Factors* of 10-K filings (2019)

(At least) ten different ways of obtaining document vectors for performing similarity comparison that have appeared in the literature.

1. Compute similarity between all firm pairs according to each method and report Pearson correlation
2. Regress pairwise similarity on firm covariates



Source: Ash, E. and S. Hansen (2023). Text Algorithms in Economics, *Annual Review of Economics*.



Source: Ash, E. and S. Hansen (2023). Text Algorithms in Economics, *Annual Review of Economics*.

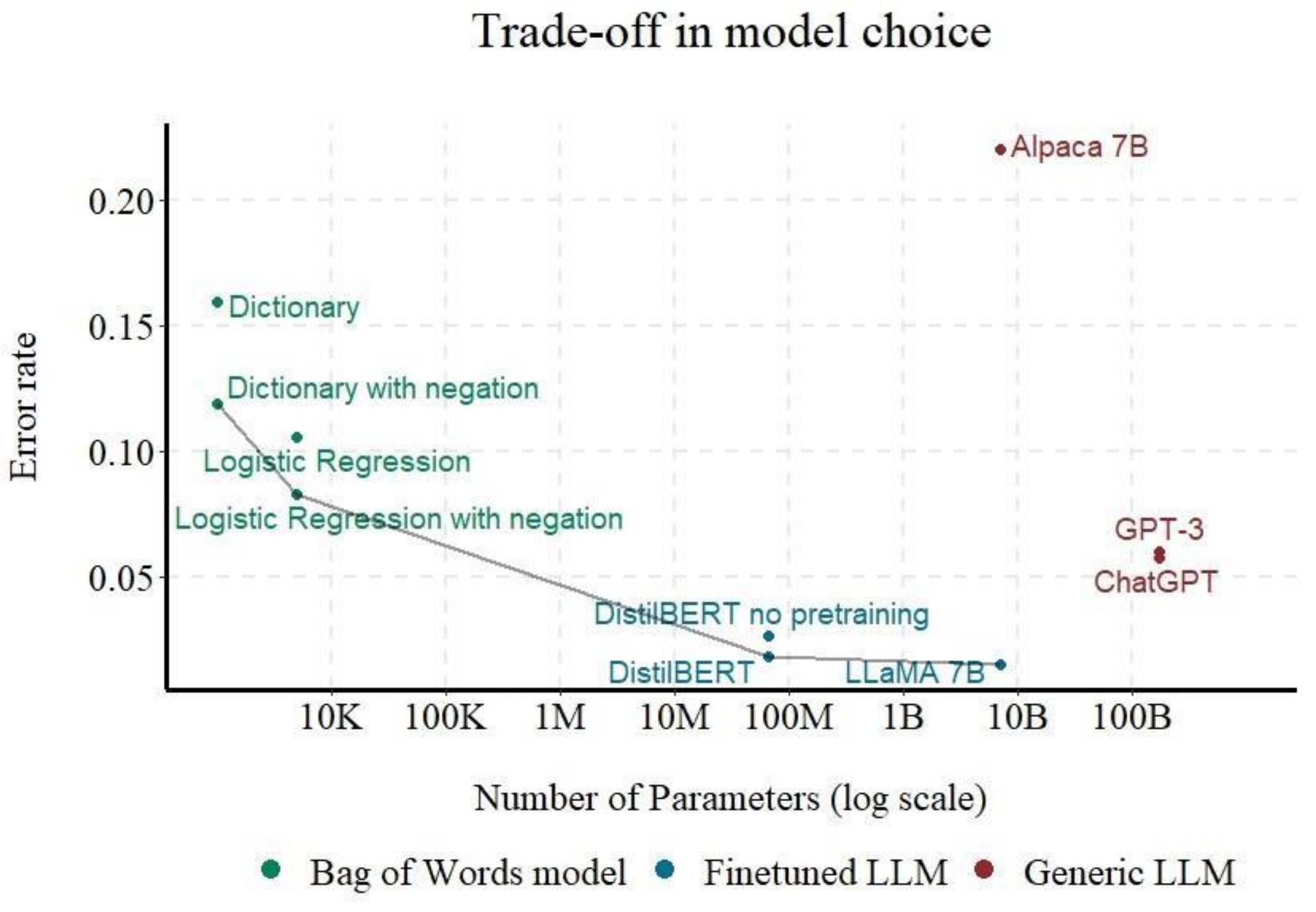
Accuracy vs. Model Size

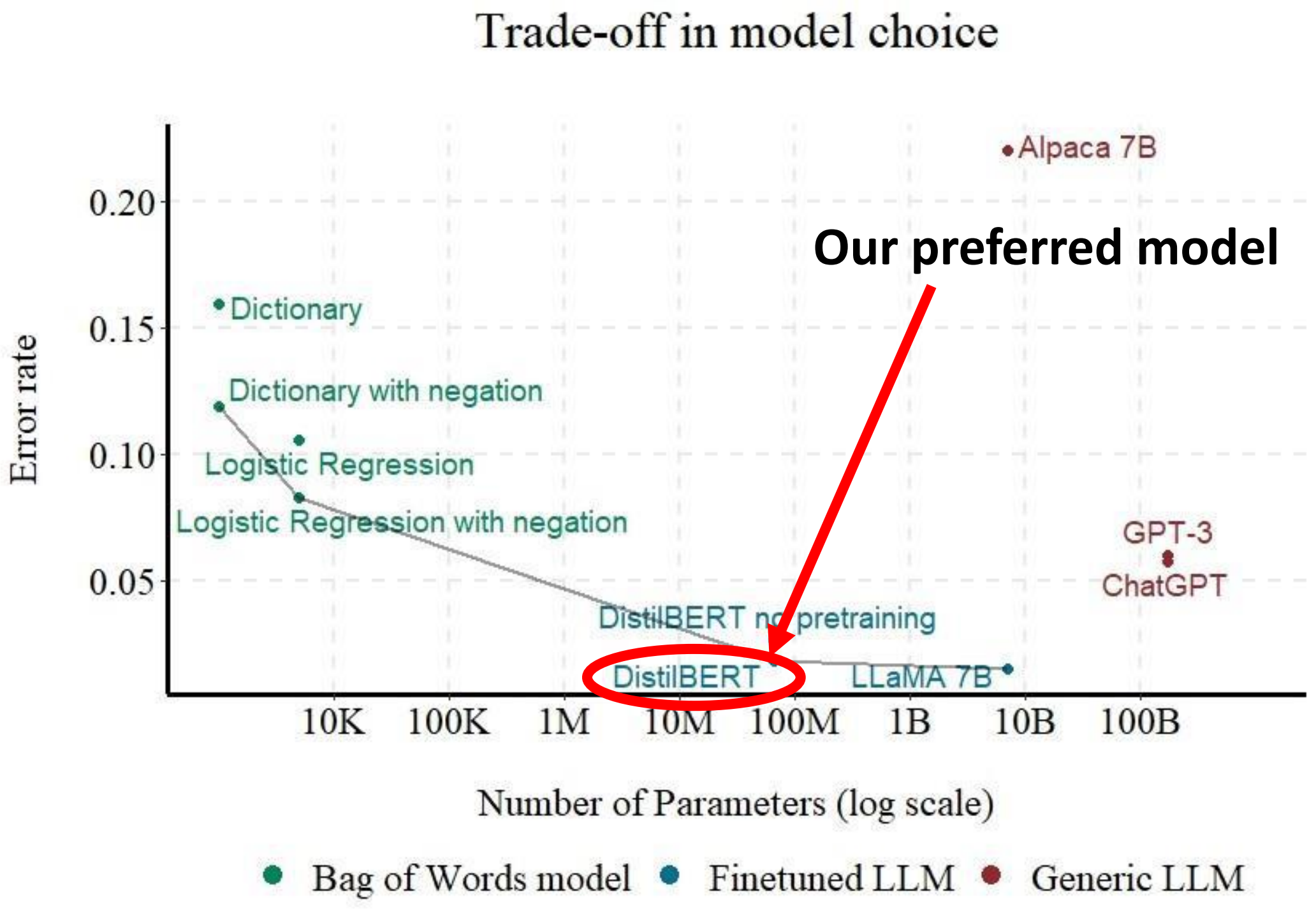
In “Remote Work…” we measure extent to which firms offer remote work using online job postings provided by Lightcast, >250m total postings

Keyword search would be easy to implement but potentially prone to error

We gather 30,000 ground truth human labels which in principle gives a validation set to guide algorithm choice

But not clear that out-of-sample goodness-of-fit is only relevant target


Trade-off in model choice
Alpaca 7B
0.20
0.15
0.10
0.05
Error rate
Dictionary
Dictionary with negation
Logistic Regression
Logistic Regression with negation
GPT-3
ChatGPT
DistilBERT no pretraining
DistilBERT
LLaMA 7B
10K
100K
1M
10M
100M
1B
10B
100B
Number of Parameters (log scale)
Bag of Words model
Finetuned LLM
Generic LLM


Trade-off in model choice
Alpaca 7B
Our preferred model
0.20
0.15
0.10
0.05
Error rate
Dictionary
Dictionary with negation
Logistic Regression
Logistic Regression with negation
GPT-3
ChatGPT
DistilBERT no pretraining
DistilBERT
LLaMA 7B
10K
100K
1M
10M
100M
1B
10B
100B
Number of Parameters (log scale)
Bag of Words model
Finetuned LLM
Generic LLM

How to Perform Inference?

As economists and policymakers, we care about inference, hypothesis testing, uncertainty quantification.

Statistical analysis is arguably becoming harder with LLMs.

Very little work so far on developing reliable inference methods for text and unstructured data.

Typically Assumed Setup

- **z**: latent variable of economic interest, e.g. policy uncertainty
- **x**: text data, e.g. newspapers
- **y**: outcome data, e.g. aggregate output

Ideal approach is to model **y** as a function of **z**

Typical approach is

(i) use **x** to create proxy measure **z'**

(ii) model **y** as a function of **z'**

When is Two-Step Approach Problematic?

In Battaglia et. al. (2024) we address this problem theoretically.

N is number of observations and C_i the total features in observation i.

Bias in distribution of regression parameters related to

$$\kappa = \sqrt{N}\, E\left[\frac{1}{C_i}\right]$$

size of dataset

variance in estimate of z_i

How Relevant is Problem?

Lightcast data

κ: 20

Nielsen Homescan data

κ: 4

10-K Business Descriptions

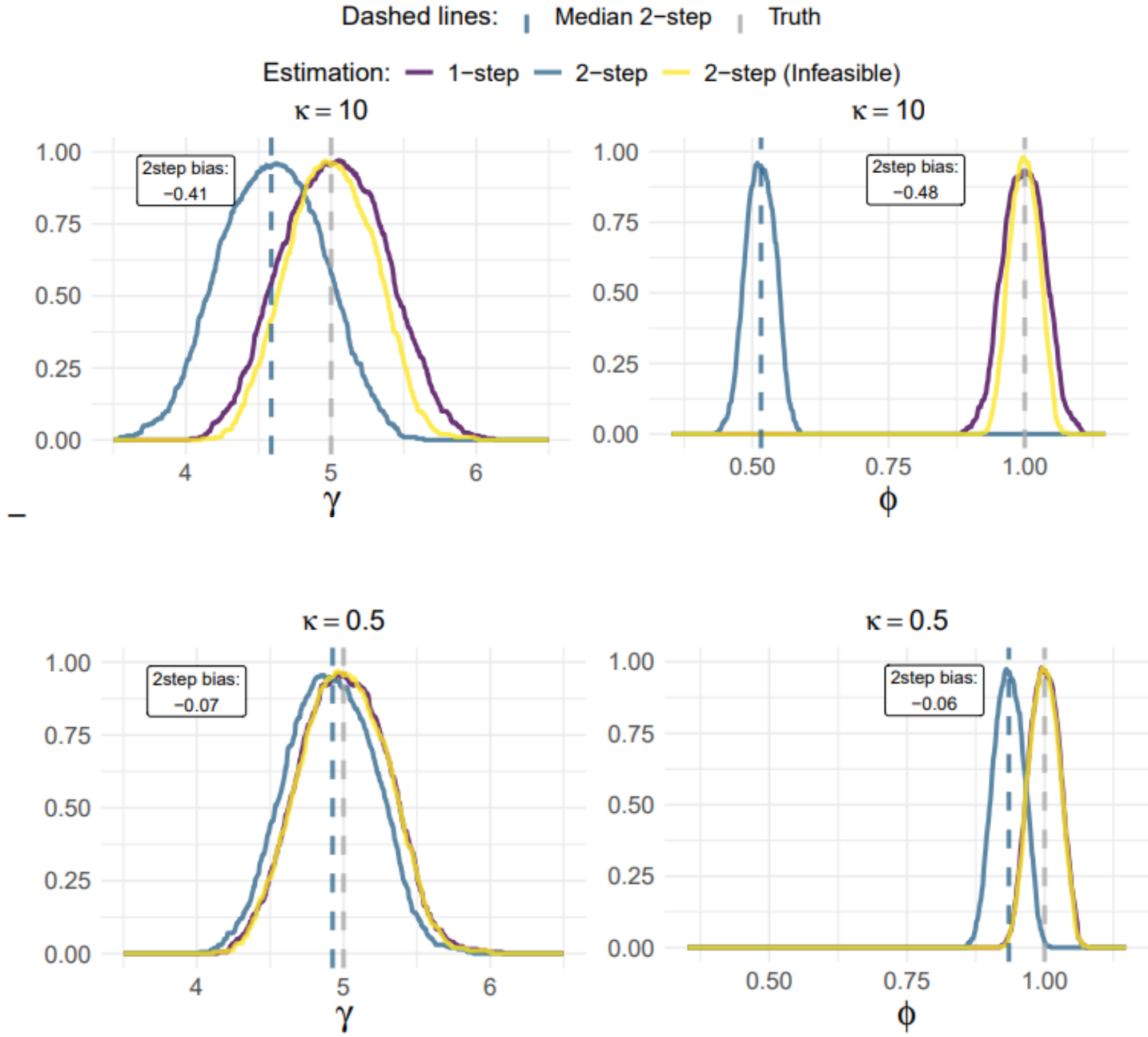
κ: 0.5

Proposed Solution

To perform robust inference, define joint likelihood over all objects.

Use Bayesian computation to draw samples from the model.

Possible at scale due to modern computational tools.


Dashed lines: Median 2−step Truth
Estimation: 1−step 2−step 2−step (Infeasible)
κ = 10
2step bias: −0.41
γ
κ = 10
2step bias: −0.48
ϕ
κ = 0.5
2step bias: −0.07
γ
κ = 0.5
2step bias: −0.06
ϕ

Conclusion

Text data has unlocked new measures in many fields of economics

This young field naturally has scope to mature by:

- Establishing clearer norms surrounding choice of model
- Establishing more consistent benchmarking exercises
- Addressing challenge of statistical inference