

# Precise Model Benchmarking with Only a Few Observations

**Riccardo Fogliato\***

Amazon Web Services  
fogliato@amazon.com

**Nil-Jana Akpinar**

Amazon Web Services  
nakpinar@amazon.com

**Pratik Patil**

University of California, Berkeley  
pratikpatil@berkeley.edu

**Mathew Monfort**

Amazon Web Services  
monfortm@amazon.com

## Abstract

How can we precisely estimate a large language model’s (LLM) accuracy on questions belonging to a specific topic within a larger question-answering dataset? The standard direct estimator, which averages the model’s accuracy on the questions in each subgroup, may exhibit high variance for subgroups (topics) with small sample sizes. Synthetic regression modeling, which leverages the model’s accuracy on questions about other topics, may yield biased estimates that are too unreliable for large subgroups. We prescribe a simple yet effective solution: an empirical Bayes (EB) estimator that balances direct and regression estimates for each subgroup separately, improving the precision of subgroup-level estimates of model performance. Our experiments on multiple datasets show that this approach consistently provides more precise estimates of the LLM performance compared to the direct and regression approaches, achieving substantial reductions in the mean squared error. Confidence intervals for EB estimates also have near-nominal coverage and are narrower compared to those for the direct estimator. Additional experiments on tabular and vision data validate the benefits of this EB approach.

## 1 Introduction

Accurate evaluation of large language models (LLMs) is crucial for identifying their strengths and weaknesses. While broad topics like math or history often have lots of data available for testing, specific topics may lack suitable test data. For example, we may not be able to collect queries for niche subjects, such as a particular historical event, legal terms from specific jurisdictions, rare medical conditions, or regional dialects. Consequently, we may fail to reliably assess the LLM’s understanding of these topics. When gathering more data is



Figure 1: **Estimates of LLM accuracy and their 95% confidence intervals** for predictions made by Gemma-2b across various subgroups on a subset of HellaSwag. The empirical Bayes estimates have precision similar to the direct estimator for some subgroups (e.g., clean and jerk) and higher precision for others (e.g., high jump). This approach also provides tighter confidence intervals for the estimates; e.g., see running and sports topics.

not an option, practitioners must rely on existing datasets to measure the model’s performance.

To illustrate this problem, consider a randomly sampled subset of HellaSwag (Zellers et al., 2019), a dataset with multiple-choice questions on various topics or subgroups; e.g., see Figure 1. Our goal is to precisely estimate the accuracy of the answers given by the LLM in each subgroup using only the limited data available. Traditionally, we estimate model performance via a direct estimator (DT), which computes the average accuracy in each subgroup separately. However, when a subgroup has only a few questions, the DT estimates can exhibit high variance and become unreliable.

One may expect the performance of the LLM on related subgroups to be associated. Synthetic estimation aims to exploit this relationship via regression modeling (SR) (Rao and Molina, 2015). For example, we can incorporate the LLM’s knowledge of sports to aid in estimating its accuracy on questions about running. The resulting estimates will generally exhibit less variance compared to direct estimation but may be biased, and thus they may be imprecise for subgroups with large sizes.

\*Corresponding author

Empirical Bayes (EB) approaches combine DT and SR estimates, adjusting the contribution of each for every subgroup separately (Efron and Morris, 1977). In theory, they hold the promise to improve the precision of these baselines (Ignatiadis and Wager, 2019) and allow for the construction of confidence intervals (Armstrong et al., 2022); e.g., see again Figure 1. However, to the best of our knowledge, an extensive empirical evaluation of the validity of this approach for precise model benchmarking is missing.

In this work, we empirically assess the precision of EB estimates of LLM performance on subgroups (domains, tasks, topics, etc.) across multiple benchmark datasets. Our results in Section 4 show that EB estimates are consistently more precise than those of DT and SR. Their confidence intervals have good coverage and substantially smaller widths than those of DT. Additional experiments on vision and tabular data validate these findings (Appendix C).

## 2 Problem Setup

Consider a dataset  $\mathcal{D} = \{(X_g, Z_g)\}_{g=1}^G$ , where  $X_g \in \mathcal{X}$  represents the features of subgroup  $g$  and  $Z_g$  indicates the performance measure for that subgroup. For example,  $X_g$  could be the description of the topic of the questions in the subgroup, while  $Z_g$  could denote the LLM’s average empirical classification accuracy or Brier score on such questions. Our goal is to use  $\mathcal{D}$  to obtain estimates of subgroup performances  $\{\mu_g\}_{g=1}^G$  as accurate as if we had infinite data for each subgroup. Formally, we define  $\mu_g := \lim_{n_g \rightarrow \infty} Z_g$ , where  $n_g$  denotes the subgroup size. The objective is to find estimators  $\{\hat{\mu}_g\}_{g=1}^G$  that minimize the average mean squared error (MSE) across all subgroups:

$$\frac{1}{G} \sum_{g=1}^G \text{mse}(\hat{\mu}_g) = \frac{1}{G} \sum_{g=1}^G \mathbb{E}[(\hat{\mu}_g - \mu_g)^2]. \quad (1)$$

We say that estimators with lower average MSE are more *precise*. By standard arguments, the MSE for each subgroup  $g$  decomposes into bias and variance components:  $\text{mse}(\hat{\mu}_g) = \text{bias}^2(\hat{\mu}_g) + \text{var}(\hat{\mu}_g)$ , where  $\text{bias}(\hat{\mu}_g) := \mathbb{E}[\hat{\mu}_g] - \mu_g$  and  $\text{var}(\hat{\mu}_g) := \mathbb{E}[(\hat{\mu}_g - \mathbb{E}[\hat{\mu}_g])^2]$ . In the following, we consider three estimation methods.

## 3 Methods

**Direct estimator (DT).** The standard approach for estimating  $\mu_g$  is the direct estimator, where we

use  $\hat{\mu}_g = Z_g$  which is the subgroup-conditional empirical average. This estimator is unbiased and its MSE is equal to its variance, i.e.,  $\text{mse}(\hat{\mu}_g) = \text{var}(\hat{\mu}_g)$ . When the subgroup size  $n_g$  is large, this variance is typically small, resulting in a precise estimate. However, for small  $n_g$ , DT suffers from high variance, leading to less reliable estimate. To quantify the uncertainty of these estimates, we use Wilson score intervals for binomial proportions (e.g., binary accuracy) (Brown et al., 2001) and Student’s t-intervals for other continuous outcomes (e.g., F1 score).

**Synthetic regression (SR).** Synthetic regression estimators leverage information from related subgroups by learning a regression function  $f(X_g) = \mathbb{E}[Z_g | X_g]$ . This is a common approach in the small area estimation literature (Rao and Molina, 2015). The estimator is then given by  $\hat{\mu}_g = \hat{f}(X_g)$ , where  $\hat{f}$  is the fitted regression model. We use XGBoost with cross validation (Chen and Guestrin, 2016), a flexible method that reduces variance through regularization. As features, we employ the text embeddings of  $X_g$ , LLM confidence scores, as well as task- and model-specific intercepts; other approaches are possible, see Appendix A for more details. While  $\hat{f}(X_g)$  may introduce bias, the hope is that it significantly reduces variance compared to DT, especially for groups with small  $n_g$ ’s. This would result in  $\text{var}(\hat{\mu}_g) \ll \text{bias}(\hat{\mu}_g)$ . SR provides the greatest improvements in precision over the DT estimator when the regression fit is good and the subgroup metric  $Z_g$  has high variance.

**Empirical Bayes estimator (EB).** The empirical Bayes estimator combines the strengths of DT and SR. The estimator is formally given by:

$$\hat{\mu}_g = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{A}} \cdot \hat{f}(X_g) + \frac{\hat{A}}{\hat{\sigma}_g^2 + \hat{A}} \cdot Z_g, \quad (2)$$

where  $\hat{\sigma}_g^2 = \widehat{\text{var}}(Z_g)$  is an estimate of the variance of  $Z_g$ , and  $\hat{A} = \{G^{-1} \sum_{g=1}^G [(Z_g - \hat{f}(X_g))^2 - \hat{\sigma}_g^2]\}_+$ . The EB estimator dynamically balances the DT and SR estimates, aligning more closely with SR when  $\hat{f}(X_g)$  approximates  $\mu_g$  accurately and with DT otherwise. Theoretical results from Ignatiadis and Wager (2019) show that, under certain regularity conditions and parameter estimation through sample splitting, EB has lower MSE compared to DT and SR. Additionally, confidence intervals can be obtained for EB estimates, as detailed in Armstrong et al. (2022). We describe their

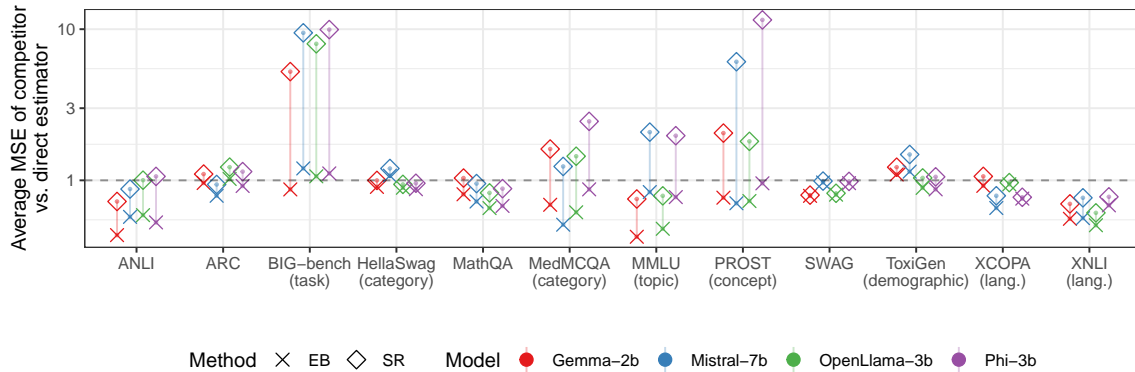


Figure 2: **Comparison of methods to estimate the accuracy of LLMs across datasets.** The plot shows the ratios of the estimates MSEs, obtained using regression (SR) and empirical Bayes (EB) methods, relative to the direct estimator (DT) for the LLM’s accuracies on LLM-task subgroups (in parenthesis when pre-defined). Lower ratio values indicate more accurate estimates compared to DT. EB consistently provides more precise estimates than both SR and DT across most evaluations.

construction in Appendix A.

## 4 Results

In this section, we present the core set of results, conveying our key takeaways. We defer to Appendix C for additional experiments, including a comparison with other baselines and other data modalities.

### 4.1 Tasks and models

We access the accuracy of LLMs across multiple-choice question-answering (MC QA) datasets, including a subset of BIG-bench (Srivastava et al., 2022), HellaSwag (Zellers et al., 2019), MedMCQA (Pal et al., 2022), MMLU (Hendrycks et al., 2020), PROST (Aroca-Ouellette et al., 2021), ToxiGen (Hartvigsen et al., 2022), XNLI (Conneau et al., 2018), and XCOPA (Ponti et al., 2020). Each dataset is analyzed independently, with subgroups defined by topics, domains, or tasks defined in the data. Additionally, we also examine ANLI (Nie et al., 2020), ARC (Clark et al., 2018), MathQA (Amini et al., 2019), and Swag (Zellers et al., 2018), which do not have pre-defined categories. On these datasets, we form subgroups through unsupervised clustering of the text embeddings via  $k$ -means, choose the number via the silhouette score, and then aggregate smaller subgroups. Lastly, we also consider tasks involving text generation included in BIG-bench, SQuAD2.0 (Rajpurkar et al., 2018), and TriviaQA (Joshi et al., 2017). The number of subgroups  $G$  varies between 40 (PROST) and more than 200 (BIG-bench).

In terms of LLMs, we use Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) Google Gemma-2b (Team et al., 2024), Microsoft Phi-3-Mini-4K-Instruct (Abdin et al., 2024), and OpenLLama-3B-v2 (Geng

and Liu, 2023). All text embeddings are generated using an off-the-shelf BERT encoder (Devlin et al., 2018).

### 4.2 Experimental setup

To mimic what a practitioner would do, we consider each dataset separately but benchmark multiple LLMs at the same time. Thus, our evaluation process is as follows: (i) *Data sampling.* Sample evaluation data from the entire dataset proportionally to subgroup sizes, ensuring that the smallest subgroup with at least 50 observations includes a minimum of  $n_g = 10$  observations in the sample. We also experiment with  $n_g = 10, 20, 50$  for all subgroups, see Appendix C.1. (ii) *Model tuning and estimation.* Tune regression models via cross-validation, estimate the subgroup metrics  $\hat{\mu}_g$  on the evaluation data using DT, SR, and EB for all LLMs, and compute the corresponding confidence intervals. (iii) *Repeated sampling and evaluation.* Repeat the process 1000 times and compute the MSE for each subgroup. “Ground truth” subgroup metrics  $\mu_g$  are estimated on the entire dataset. We compare the methods by their average MSE as in (1). A lower average MSE indicates a better method.

**EB estimation.** Our implementation of the EB approach is based on the standard cross-fitting procedure of Ignatiadis and Wager (2019) and is summarized in the following steps. (1) Estimate  $\hat{\sigma}_g^2$  for all  $g$ ’s. (2) Split the sample into two folds,  $\mathcal{D}_1$  and  $\mathcal{D}_2$ . (3) Fit SR on  $\mathcal{D}_1$ , obtain SR estimates on  $\mathcal{D}_2$  and compute  $\hat{A}$ . (4) Use these to obtain EB estimate for each  $g$  in  $\mathcal{D}_2$ . (5) Repeat the process with the folds inverted to generate the EB estimates  $\{\hat{\mu}_g\}_{g=1}^G$ .

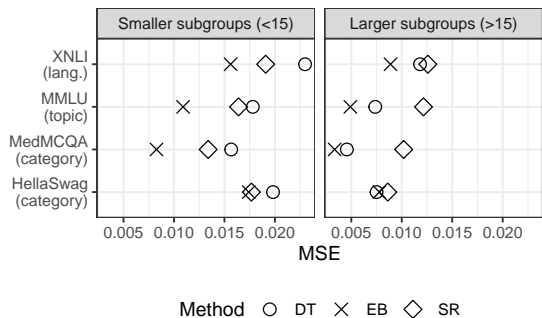


Figure 3: **Comparison of subgroup MSEs across methods.** The plot compares the MSEs across all subgroups (LLM-domain pairs) across four datasets. SR tends to perform better than DT on small subgroups but not always on larger ones. EB performs better than both on either. An  $\text{MSE} = 0.01$  means that, on average, we have  $|\hat{\mu}_g - \mu_g| = 0.1$ .

### 4.3 Results

**Accuracy estimation in MC QA.** Figure 2 gives a finegrained view of our results, showing the ratio of the MSE for the EB and SR estimators compared to the DT estimator, across various models and datasets. The EB estimator consistently performs as well as or better than the two baselines, achieving on average 20% and 30% lower MSEs than SR and DT respectively. SR generally outperforms DT, although there are exceptions, e.g., BIG-bench, where most of the subgroups are large (and therefore DT estimates are precise). On other datasets (e.g., ANLI), SR outperforms DT, but EB typically still emerges as the most precise estimator.

Figure 3 separates the MSE for larger ( $> 15$  questions) and smaller ( $\leq 15$  questions) subgroups. We observe that DT generally provides more accurate estimates than SR on the former due to small subgroup variances  $\sigma_g^2$ . In these cases, EB estimates align closely with those from DT. For the smaller subgroups, SR tends to present lower MSE than DT. In these cases, EB estimates mostly align with SR’s, resulting in improved precision compared to DT. Thus, by shrinking the estimates of SR more heavily towards those of DT’s when subgroup sizes are large (e.g., Big-bench) or when the regression fit is poor (e.g., MMLU, PROST), implying that  $\hat{A}/\hat{\sigma}^2$  is large, EB yields more precise estimates compared than the baselines.

To determine the drivers of the precision gains of SR over DT, we refit SR removing groups of features one at a time, following the LOCO (Leave Out COvariates) approach of Lei et al. (2018); we refer readers to Verdinelli and Wasserman (2023) for comparisons between this strategy and Shapely values (Lundberg, 2017). Across all datasets, we

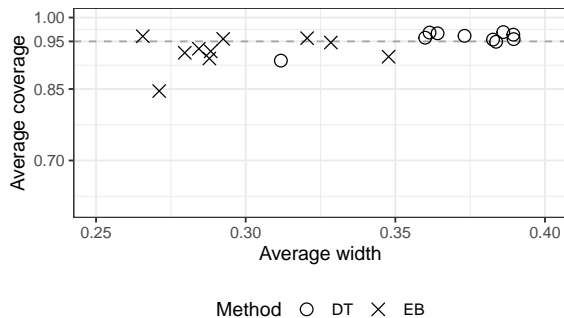


Figure 4: **Average coverage and width of 95% confidence intervals** for DT and EB estimates of LLM accuracy across datasets. EB intervals maintain high coverage and are generally narrower than those of DT.

find that removing model-specific intercepts as well as confidence scores leads to the largest increase in MSE, which we attribute to variations in the accuracy of the LLMs. The removal of embeddings and task-related features do not appear to be strongly associated with the precision of the estimates.

Overall, we find that due to strong regularization, SR estimates vary across LLMs but remain consistent across tasks for the same LLM. This strategy is effective only when subgroup performance is similar within a given LLM, aligning with our observation that most features are not highly informative. Considering also the sizes of the subgroups—typically large in BIG-bench and small in MMLU—this explains the performance differences between SR and DT. Lastly, benchmarking LLMs separately only slightly decreases the efficiency of EB, which still outperforms DT. Thus, even if only one LLM is evaluated, EB should be used.

#### Coverage and width of confidence intervals.

We evaluate confidence intervals in terms of their coverage (i.e., the frequency with which they capture  $\mu_g$ ) and their width, averaged across all subgroups. Since EB estimates are more precise than SR, we focus on comparing DT and EB. Ideally, confidence intervals should achieve nominal coverage and be as narrow as possible. As shown in Figure 4, DT intervals consistently achieve nominal coverage (here 95%), while EB intervals also maintain high coverage, with most exceeding  $\geq 90\%$ . However, the average width of EB intervals is on average 20% smaller than those of DT, indicating that EB estimates come with tighter confidence intervals.

**Estimation for general tasks.** Finally, we extend our comparison of estimation methods to tasks beyond MC QA. Table 1 summarizes the



Data/MSE $\times 10^3$	DT	SR	EB
BIG-bench ( <i>exact match</i> )	<b>1.0</b>	19.5	1.6
SQuAD2.0 ( <i>F1 score</i> )	246.0	178.0	<b>109.5</b>
Trivia QA ( <i>exact match</i> )	12.2	12.8	<b>9.6</b>

Table 1: **Estimation method comparison on general tasks.** The table compares the average MSEs of the estimates of all LLM performances across different datasets (with respective evaluation metrics) by the different methods. EB consistently achieves similar or lower MSE than DT and SR.

average MSE of each method on the BIG-bench, SQuAD2.0, and Trivia QA datasets. While DT outperforms SR on BIG-bench, SR outperforms DT on SQuAD2.0. Their performance is similar on TriviaQA. EB remains the best-performing estimator on two datasets and its performance is close to that of DT on BIG-bench.

## 5 Discussion

Our investigation contributes to a growing body of work on model evaluation in presence of resource constraints (Fogliato et al., 2024; Zrnic and Candès, 2024; Herlihy et al., 2024). Our results show that an EB estimator, which combines DT and SR estimates, consistently provides more precise estimates of subgroup-level model performances compared to DT and SR alone. When subgroups are large, the estimates of EB align with those of DT. The uncertainty of EB estimates can also be effectively quantified via confidence intervals, which achieve near-nominal coverage and are consistently tighter than those derived for DT estimates (although see limitations in Appendix B). Overall, EB proves to be a simple and useful method for model evaluation when estimating performance for lots of subgroups with (many or) few observations.

## References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. **MathQA: Towards interpretable math word problem solving with operation-based formalisms.** In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. **Machine bias: There’s software used across the country to predict future criminals. and it’s biased against blacks.** *ProPublica*.

Timothy B Armstrong, Michal Kolesár, and Mikkel Plagborg-Møller. 2022. Robust empirical bayes confidence intervals. *Econometrica*, 90(6):2567–2602.

Stéphane Aroca-Ouellette, Cory Paik, Alessandro Roncone, and Katharina Kann. 2021. Prost: Physical reasoning of objects through space and time. *arXiv preprint arXiv:2106.03634*.

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32.

Dillon Bowen. 2022. Multiple inference: A python package for comparing multiple parameters. *Journal of Open Source Software*, 7(75):4492.

Lawrence D Brown, T Tony Cai, and Anirban DasGupta. 2001. Interval estimation for a binomial proportion. *Statistical science*, 16(2):101–133.

David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. 2023. **CLAIR: Evaluating image captions with large language models.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646, Singapore. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. **Microsoft COCO captions: Data collection and evaluation server.** *CoRR*, abs/1504.00325.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. 2017. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

- Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Paulo Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance.
- Li Deng. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. **Retiring adult: New datasets for fair machine learning**. In *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490. Curran Associates, Inc.
- Bradley Efron and Carl Morris. 1975. Data analysis using stein’s estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319.
- Bradley Efron and Carl Morris. 1977. Stein’s paradox in statistics. *Scientific American*, 236(5):119–127.
- Dugas Emma, Jorge Jared, and Will Cukierski. 2015. **Diabetic retinopathy detection**.
- M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. <http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>
- Robert E Fay III and Roger A Herriot. 1979. Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277.
- Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE.
- Riccardo Fogliato, Pratik Patil, Mathew Monfort, and Pietro Perona. 2024. A framework for efficient model evaluation through stratification, sampling, and estimation. *arXiv preprint arXiv:2406.07320*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. **A framework for few-shot language model evaluation**.
- Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*.
- Xinyang Geng and Hao Liu. 2023. **Openllama: An open reproduction of llama**.
- M Ghosh and JNK Rao. 1994. Small area estimation: An appraisal. *Statistical Science*, 9(1):55–76.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.
- Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
- Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021b. Natural adversarial examples. *CVPR*.
- Christine Herlihy, Kimberly Truong, Alexandra Chouldechova, and Miroslav Dudik. 2024. A structured regression approach for evaluating model performance across intersectional subgroups. *arXiv preprint arXiv:2401.14893*.
- Nikolaos Ignatiadis and Stefan Wager. 2019. Covariate-powered empirical bayes estimation. *Advances in Neural Information Processing Systems*, 32.
- Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. **Openclip**.

- William James and Charles Stein. 1992. Estimation with quadratic loss. In *Breakthroughs in statistics: Foundations and basic theory*, pages 443–460. Springer.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia.
- Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.
- LAION AI. 2024. Clip benchmark. [https://github.com/LAION-AI/CLIP\\_benchmark](https://github.com/LAION-AI/CLIP_benchmark).
- Nan M Laird and Thomas A Louis. 1987. Empirical bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82(399):739–750.
- Yann LeCun, Fu Jie Huang, and Leon Bottou. 2004. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104. IEEE.
- Jing Lei, Max G’Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.
- Scott Lundberg. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.
- Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.
- Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.
- Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.
- Pratik Patil, Arun Kumar Kuchibhotla, Yuting Wei, and Alessandro Rinaldo. 2022. Mitigating multiple descents: A model-agnostic framework for risk monotonization. *arXiv preprint arXiv:2205.12937*.
- Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal common-sense reasoning. *arXiv preprint arXiv:2005.00333*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don’t know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

- John NK Rao and Isabel Molina. 2015. *Small area estimation*. John Wiley & Sons.
- Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.
- Evan TR Rosenman, Guillaume Basse, Art B Owen, and Mike Baiocchi. 2023. Combining observational and experimental datasets using shrinkage estimators. *Biometrics*, 79(4):2961–2973.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. [ImageNet Large Scale Visual Recognition Challenge](#). *International Journal of Computer Vision (IJCV)*, 115(3):211–252.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. [LAION-5b: An open large-scale dataset for training next generation image-text models](#). In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. 2011. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.
- Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. 2018. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II 11*, pages 210–218. Springer.
- Isabella Verdinelli and Larry Wasserman. 2023. Feature importance: A closer look at shapley values and loco. *arXiv preprint arXiv:2303.05981*.
- Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518.
- J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. 2010. [Sun database: Large-scale scene recognition from abbey to zoo](#). In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492.
- Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.
- Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruysen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. 2020. [The visual task adaptation benchmark](#).
- Tijana Zrnica and Emmanuel J. Candès. 2024. Active statistical inference. *arXiv preprint arXiv:2403.03208*.



This serves as an appendix to the paper “Precise Model Benchmarking with Only a Few Observations.” The content of the appendix is organized as follows.

## Organization

- In Appendix A, we provide additional details about the methods used in our paper. In particular, we discuss in detail the following methods:
  - Appendix A.1: synthetic regression
  - Appendix A.2: empirical Bayes
- In Appendix B, we discuss the limitations of the methods.
- In Appendix C, we present additional experiments to further validate our results. In particular, we show results on the following data types:
  - Appendix C.1: additional LLMs
  - Appendix C.2: computer vision tasks
  - Appendix C.3: image captioning task
  - Appendix C.4: tabular data

## A Additional Details on Methods

In this section, we provide additional details on the methods discussed in the main text and discuss other baselines. While the main body of the manuscript focuses on an essential comparison, here we also present methods proposed by previous work.

### A.1 Synthetic regression (SR)

The term “synthetic” is taken from the small area literature, where it is used to indicate estimators that leverage information from large subgroups about a target quantity to derive estimates on smaller subgroups under the assumption that these subgroups share similar characteristics associated with the target quantity (Ghosh and Rao, 1994; Rao and Molina, 2015).

**Implementation.** In principle, the synthetic regression approach can utilize any family of predictive models (e.g., boosted trees, linear regression, neural nets). We use a XGBoost regressor predictor to predict  $Z_g$ , tuning the number of trees and the depth via 2-fold cross-validation, stratifying by the task type. In the covariates  $\tilde{X}_g$ , we include (i) the values of the embeddings averaged across all observations in the subgroup, (ii) the average LLM confidence scores, (iii) intercepts for each model

being evaluated and (iv) for each task. Other feature construction methods are also possible and might lead to better results; we focus on these regression features because they are intuitive and straightforward to obtain.

**Structured regression.** A special case of SR is *structured regression*, which has recently been proposed by Herlihy et al. (2024). This approach involves fitting a Lasso regression model (Tibshirani, 1996) to minimize the following loss function:

$$\arg \min_{\beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}^p} \sum_{g=1}^G \frac{1}{\sigma_g^2} (\beta^\top \tilde{X}_g - Z_g)^2 + \lambda \sum_{i=1}^p |\beta_i|, \quad (3)$$

where  $\beta = (\beta_0, \dots, \beta_p)^\top \in \mathbb{R}^p$ . We include subgroup-level intercepts in the feature set  $\tilde{X}_g$  (e.g., on for each LLM-task pair), along with other explanatory features as in SR (e.g., the embeddings). We tune  $\lambda$  via cross-validation and estimate  $\sigma_g^2$  using the plug-in estimator, e.g.,  $Z_g(1 - Z_g)/n_g$  for accuracy metrics. Note that if the number of features outnumbers the number of observations as it occurs in our experiments, we should expect that the predictions of this approach will converge to  $Z_g$  as  $\lambda \rightarrow 0^+$ . This is because the estimator in (3) “interpolates” the training data and, among all such interpolators, it has the minimum  $\ell_1$ -norm; see, e.g., Patil et al. (2022).

**Methods comparison.** We distinguish SR from structured regression to emphasize the fact that any modeling approach and feature construction can be used in the former. For example, we arbitrarily choose not to include subgroup-level intercepts in our regression model. Structured regression represents an alternative to our EB approach, where shrinkage is integrated directly into SR instead of explicitly through the  $\hat{A}$  and  $\sigma_g^2$  parameters. Overall, we find that EB is more flexible because practitioners can use any predictive model and feature set, making it easier to implement. In practice, we will see that, when both approaches use Lasso regression, their performance is similar.

### A.2 Empirical Bayes (EB)

The empirical Bayes estimator in our work is motivated by the following generative model (Armstrong et al., 2022): For  $1 \leq g \leq G$ , assume that

$$\mu_g \mid X_g, \sigma_g^2 \sim \mathcal{N}(f(X_g), A), \quad (4)$$

$$Z_g \mid \mu_g, X_g, \sigma_g^2 \sim \mathcal{N}(\mu_g, \sigma_g^2) \quad (5)$$

where  $A > 0$ . Estimator (2) is derived from the posterior distribution of this data-generation mechanism. We recall it here:

$$\hat{\mu}_g = \hat{f}(X_g) + \frac{\hat{A}}{\hat{\sigma}_g^2 + \hat{A}} \cdot (Z_g - \hat{f}(X_g)), \quad (6)$$

where

$$\hat{A} = \left( \frac{1}{G} \sum_{g=1}^G \varepsilon_g^2 - \hat{\sigma}_g^2 \right)_+. \quad (7)$$

Here,  $\varepsilon_g = Z_g - \hat{f}(X_g)$ ,  $\hat{\sigma}_g^2$  is estimated as in the SR approach, and  $(x)_+$  denotes the positive part of a real number  $x$ .

**Boundary behavior.** Observe the following two extreme cases:

- If  $\hat{\sigma}_g^2 \ll \hat{A}$ , EB estimates will be close to DT. This case can occur when  $n_g$  is large or SR does not explain well the  $\mu_g$ 's.
- If  $\hat{\sigma}_g^2 \gg \hat{A}$ , EB estimates align with SR. This case can occur when  $Z_g$  has high variance.

**Confidence intervals.** There are several methods for constructing confidence intervals for EB estimates (Laird and Louis, 1987; Rosenman et al., 2023). We choose the approach of Armstrong et al. (2022), which is robust to violations of the assumption in (4). More specifically, the confidence intervals are valid even when (4) is replaced by the following assumption:

$$\mathbb{E}[(\mu_g - f(X_g))^2 | X_g, \sigma_g^2] = A. \quad (8)$$

Their construction is fairly technical but has been implemented in existing software (Bowen, 2022). Briefly, let

$$\hat{\kappa} = \sum_{g=1}^G \frac{\varepsilon_g^4 - 6\hat{\sigma}_g^2 \varepsilon_g^2 + 3\hat{\sigma}_g^4}{\hat{A}^2}. \quad (9)$$

Then the  $1 - \alpha$  confidence interval for  $\mu_g$  is given by:

$$\hat{\mu}_g \pm \text{cva}_\alpha(\hat{\sigma}_g^2/\hat{A}, \hat{\kappa}) \cdot \frac{\hat{A}}{\hat{\sigma}_g^2 + \hat{A}} \cdot \hat{\sigma}_g, \quad (10)$$

where the so-called critical value  $\text{cva}_\alpha$  replaces the traditional percentile of the standard normal distribution in the typical intervals to account for the bias in the shrinkage (Armstrong et al., 2022). Differently from the standard intervals that we construct for DT estimates, these intervals have an ‘‘average

coverage’’ property: While DT intervals cover each parameter with likelihood at least  $1 - \alpha$ , the average coverage level of EB intervals across all  $\mu_g$ 's is at least  $1 - \alpha$ .

**More background on composite estimators.** Estimators of the form (2), which combine DT with other estimates, are more broadly known as composite estimators (Rao and Molina, 2015). This class includes the classical James-Stein estimator (James and Stein, 1992), which shrinks all  $Z_g$  values towards the same chosen value, which is generally taken to be the global mean, e.g., see Section 2 in Efron and Morris (1975). Another estimator commonly used in small area estimation is the Fay-Herriot model, which assumes that  $f(X)$  is linear (Fay III and Herriot, 1979). The version of EB that we study allows for arbitrary SR models and thus covers these cases. We expect the James-Stein estimator to perform similarly in cases where the regression employs strong regularization.

## B Limitations of Methods

Empirical Bayes methods are built on stronger assumptions than the standard direct estimator. These assumptions can allow us to increase the precision of the estimates and our experiments have shown that in general this occurs. However, this is not guaranteed and, even when it occurs, it may not always be desirable. For example, if one is interested in estimating only a subset of the  $\mu_g$ 's, then they should not seek to minimize (1). Another critical observation is the nature of the confidence intervals for EB estimates that we discuss. While standard intervals for DT are constructed to cover each  $\mu_g$  with a given likelihood, EB intervals only have an average coverage guarantee. This should be noted when we aim to conduct hypothesis testing on the  $\mu_g$ 's using EB estimates. Practitioners that misunderstand this point might risk making erroneous inference and draw unsupported conclusions. In addition, the efficacy of EB over DT also hinges on how well the regression models  $\mathbb{E}[Z_g | X_g]$ . Therefore, feature construction and model selection are crucial. Our research has begun to explore these aspects, but further investigation is needed.

## C Additional Experiments

In this section, we provide additional details on our main set of results and additional experiments to evaluate the performance of the methods.

Size/Rel. eff.	MSE SR / DT	MSE EB / DT
$n_g = 10$	0.90	0.81
$n_g = 20$	1.03	0.84
$n_g = 50$	1.52	0.86

Table 2: **Comparison of estimation methods on LLM MC QA tasks with subgroups of equal size.** The table compares the median over datasets of the relative efficiency (rel. eff.) of DT with respect to SR and EB, namely the ratio of the average MSE of SR or EB divided by the average MSE of DT when all subgroups have size  $n_g$ . As  $n_g$  increases, efficiency gains over DT decrease.

## C.1 Additional experiments on LLMs

To generate the LLM confidence scores and associated predictions in all our experiments, we use the code in the lm-evaluation-harness repository by Eleuther AI (Gao et al., 2023). In this section, we review the results of additional experiments, including:

- An experiment where all subgroups have exactly  $n_g \in \{10, 20, 50\}$  observations (Appendix C.1.1), including an analysis of how EB balances DT and SR estimates
- A comparison between EB and structured regression

### C.1.1 Results on all subgroups with equal size

In this benchmarking exercise, we equalize the number of observations drawn from equal subgroups, setting either  $n_g = 10$ ,  $n_g = 20$ , or  $n_g = 50$ . We drop all subgroups with less than  $4n_g$  observations. Thus, the experimental setup follows Section 4.2 with the exception of (i).

The results from this experiment are summarized in Table 2, which shows the mean relative efficiency of DT vs. SR or EB across datasets, which is defined as the ratio of the average MSE of SR or EB estimates over the DT one. We observe that, as the size of the subgroups increases, the gap in the precision of SR and DT estimates shrinks. This occurs because DT estimates become more accurate. EB shows a similar trend that is nonetheless less remarked: Even for  $n_g = 50$ , EB estimates have lower MSEs than those of DT. For this number of observations, 95% confidence intervals for DT are fairly large (e.g.,  $\hat{\mu}_g \pm 0.08$  for  $\mu_g = 0.1$ ). Consequently, the gain in precision can result in considerable savings for practitioners.

Setting/Rel. eff.	EB	Str. Reg.
$\min_{g \in [G]} n_g = 10$ , prop.	0.62 (0.14)	0.68 (0.34)
$n_g = 20$ , equal	0.64 (0.15)	0.73 (0.47)

Table 3: **Estimation method comparison on LLM MC QA tasks of EB with other baselines.** The table compares the mean (standard deviation in parentheses) over datasets of the relative efficiency (rel. eff.) of DT with respect to three methods: EB and structured regression.

## C.1.2 Comparison with structured regression

We assess the precision of EB estimates against those of the structured regression estimator described in Appendix A. Table 3 shows the corresponding results, for both the setting where subgroup sizes are proportional to the original data ( $n_g = 10$  for the smallest group) and where all  $n_g = 20$  and all subgroups have the same size. We find that EB beats structured regression across almost all datasets. The relative efficiency of their estimates, however, reveals that the methods have comparable efficiency. Thus, practitioners should choose the method that is easiest to implement.

## C.2 Experiments on computer vision tasks

We explore the performance of the methods on a series of computer vision classification tasks. We consider most of the datasets and tasks included in the LAION CLIP benchmark (LAION AI, 2024), including Caltech 101 (Fei-Fei et al., 2004), Stanford Cars (Krause et al., 2013), CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), CLEVR (distance and count) (Johnson et al., 2017), Describable Text Features (Cimpoi et al., 2014), DR Detection (Emma et al., 2015), DMLab Frames (Zhai et al., 2020), EuroSAT (Helber et al., 2019), FGVC aircraft (Maji et al., 2013), Oxford 102 Flower (Nilsback and Zisserman, 2008), GTSRB (Stallkamp et al., 2011), ImageNet-A (Hendrycks et al., 2021b), ImageNet-R (Hendrycks et al., 2021a), ImageNet-1K (Russakovsky et al., 2015), ImageNet Sketch (Wang et al., 2019), ImageNetV2 (Recht et al., 2019), KITTI Distance (Geiger et al., 2013), MNIST (Deng, 2012), ObjectNet (Barbu et al., 2019), Oxford-IIIT Pet (Parkhi et al., 2012), PASCAL VOC 2007 (Everingham et al., 2007), PCam (Veeling et al., 2018), Rendered SST-2 (Socher et al., 2013), NWPU-RESISC45 (Cheng et al., 2017), SmallNorb (Azimuth and Elevation) (LeCun et al., 2004), STL-10 (Coates et al., 2011), SUN397 (Xiao et al., 2010), Street View House Numbers (Netzer et al., 2011).

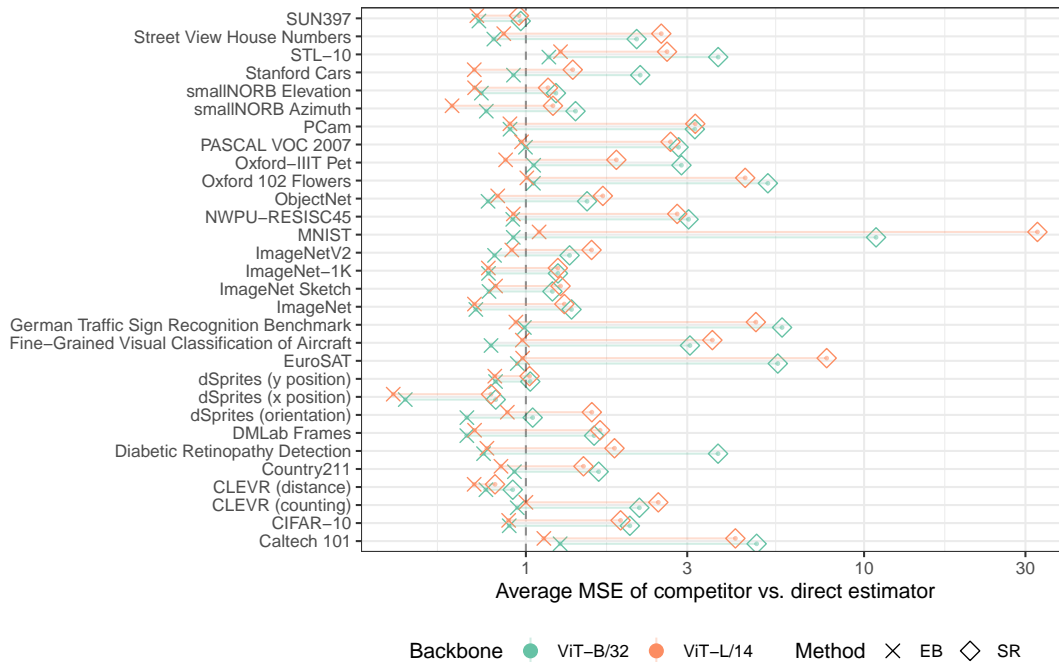


Figure 5: **Comparison of methods across datasets to estimate CLIP’s zero-shot accuracies on subgroups of classification tasks.** See the full list of tasks in Appendix C.2. The observations correspond to the ratio between the average MSE of SR or EB over the average MSE of DT estimates. EB yields more precise estimates than SR and DT across most datasets and models.

We experiment with predictions generated by CLIP models that employ ViT-B/32 and ViT-L/14 as vision backbones (Ilharco et al., 2021; Schuhmann et al., 2022; Radford et al., 2021) in the zero-shot setting. We form task subgroups via  $k$ -means as in Section 4.1 and sample  $n_g = 20$  observations from all subgroups; we found that sampling proportionally led to only a couple of small groups and many large ones. Our goal is to estimate the accuracy of the models on the subgroups. The experimental evaluation setup is analogous to what we described in Section 4.2. The setup of SR also remains unchanged, with the exception that we run the model on the ViT-B-32’s image embeddings instead of the text embeddings.

Figure 5 shows the results of this set of experiments. We find that SR beats DT on some datasets but severely underperforms on others. This is explained by the fact that (most often) the SR estimates are heavily regularized and predictions are close across subgroups. As we saw in the case of LLMs, this strategy improves over DT on datasets where subgroup performances are similar (e.g., dSprites (x position)) yet it miserably fails when subgroup performance are far apart (e.g., in MNIST). In these cases, we observe that the EB estimates align closely those from DT, as a result of a large ratio  $\hat{A}/\hat{\sigma}^2$ . Despite the overall under-

performance of SR, we find that EB still performs similarly to DT and always better than SR.

### C.3 Experiment on image captioning

We assess the performance of EB on evaluating a BLIP (Li et al., 2022) image captioning model on a test split (Karpathy and Fei-Fei, 2015) of 5000 images from the COCO Captions dataset (Lin et al., 2014; Chen et al., 2015), where each of the test images includes 5 reference captions. We use the BLIP model to generate a new caption for each image in the test set and then evaluate the results using the CLAIR score (Chan et al., 2023). This score uses an LLM (in our experiments we use a Mistral 7B model (Jiang et al., 2023)) to predict a similarity score between the generated caption and a set of reference captions. For each test image, we average the CLAIR score over the set of reference captions to obtain a single score per image.

To evaluate our approach in this setting, we form subgroups by first generating image embeddings for each image using CLIP (Radford et al., 2021) and clustering the embeddings via  $k$ -means where the silhouette score was used to choose the number of clusters (about 40 in our test case) as in the previous experiments. The rest of the experimental setup also follows Section 4.2.

We find that both SR and EB substantially out-



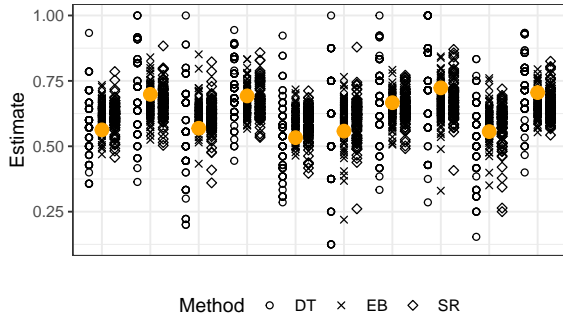


Figure 6: **Subgroup estimates of CLAIR scores across 10 subgroups in VQA data.** The dots in orange correspond to ground truth values, while those in black indicate the estimates obtained over 100 random draws from the subgroups with the different approaches.

perform DT, achieving average MSEs that are 70% and 80% lower than those of DT, respectively. Figure 6 shows the estimates across 10 randomly sampled subgroups in the data. We observe that, due to the small subgroup sizes in the evaluation data, the accuracy estimates of DT have high variance. Both those of DT and of EB have significantly less variance and, as our empirical results show, also happen to have little bias for most of the subgroups.

#### C.4 Experiments on tabular data

We present results for the subgroup performance estimation methods on a host of binary classification tasks commonly used in the algorithmic fairness literature. This includes five prediction tasks on 2018 ACS folkstable data from New York (Ding et al., 2021), the COMPAS dataset (Angwin et al., 2016), and student performance datasets in math and Portuguese (Cortez and Silva, 2008). Similar to Herlihy et al. (2024), we define subgroups by intersections of protected attributes like race or ethnicity, age, and sex. We then retain only those with at least 20 observations for subgroup performance estimation.

Predictions on tabular datasets are generated using XGBoost (Chen and Guestrin, 2016). Each model is trained with 3-fold cross-validation on 50% of the respective dataset. The other 50% of the examples are retained for our evaluation set using the same setup of Section 3.

Figure 7 shows the ratio of the MSE of EB and SR estimates to the DT estimates across all tabular datasets. Similar to our observations in the LLM context, the EB estimator provides estimates that are just as precise or more precise than SR and DT estimators for six out of the eight prediction tasks. The EB estimator’s advantage over SR is

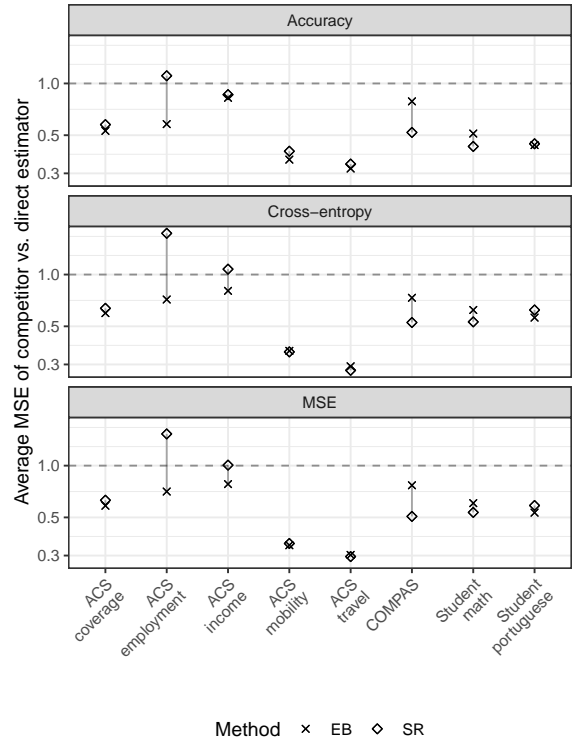


Figure 7: **Comparison of methods across tabular datasets.** The plot shows the ratio of the MSE of the estimates of the LLM metric on the subgroups obtained through SR and EB (the competitors) over the MSE of those generated by DT. Lower values indicate more precise estimates compared to DT. As metrics, we consider accuracy, cross-entropy, and MSE or Brier score.

especially pronounced for the ACS Employment and Income datasets for which SR struggles to outperform the DT baseline. EB and SR performances are virtually identical for the ACS mobility and travel tasks, implying that SR explains little to no additional variance in these cases. In the context of the COMPAS and Student Math datasets, the SR estimator provides better estimates than the EB estimator. Both of these datasets have less than 10 distinct subgroups which may lead to poor fits in computing  $\hat{f}(X)$  and  $\hat{A}$ . Nevertheless, both SR and EB estimators consistently outperform the DT baseline.

We thus conclude that, similar to the results in the main body of the paper, the EB estimator for tabular data provides more precise subgroup performance estimates than DT and also SR when there are enough distinct subgroups.