Can BERT Tell Me What GPT-3.5 Will Say? An Analysis of Predictive Correlations across Language Models

Rajan P. Vivek
Department of Computer Science
Stanford University
rvivek@stanford.edu

Abstract

Vivek et al. [2023] recently proposed Anchor Points, a method of estimating instance-level language model predictions on a large number of instances by exploiting predictive correlations shared by language models. While this method demonstrates strong empirical performance, it is currently unclear how well anchor points will generalize from the source models to which they are fit. In this work, we show that anchor points can generalize across large predictive gaps between source and target models. This generalization deteriorates as a function of the size of the predictive gap when conditioned on a set of source models. We also show that anchor points can be transferred between model families, e.g. from finetuned BERT models to prompted GPT-3.5 models, to varying extents. Finally, we call for researchers to release their model predictions at the instance level to faciliate further research into the phenomenon of inter-instance correlations across language models.

1 Introduction

Running inference on today's most powerful language models is expensive. In order to benchmark a model on a new task, find prompts that elicit the best performance, or obtain machine-generated labels for a dataset (Taori et al. [2023], Smith et al. [2022]), researchers and practitioners must pay high compute or API costs to run inference on many samples. Vivek et al. [2023] recently proposed Anchor Points, a method of selecting a subset of instances from a given dataset such that evaluating a model on the subset provides sufficient information to estimate what the model will predict on the remainder of the dataset. This technique relies on the fact that a language model's output on one sample often linearly correlates with its output on other samples in a manner that is consistent across models, i.e. $f_n(x_2) \approx w_{1,2} f_n(x_1) + b_{1,2}$, where $w_{1,2}$ and $b_{1,2}$ can be found by fitting a trend line through the predictions of other models $\{(f_1(x_1), f_1(x_2)), ...(f_{n-1}(x_1), f_{n-1}(x_2))\}$. The authors propose a simple linear model that fits to the predictions of a base set of "source" models on a given dataset. Estimating the predictions of a new "target" model (i.e. a model whose predictions we wish to know) on the entire dataset then requires evaluating the model on only a small subset of informative points.[1].

While Vivek et al. [2023] achieve impressive performance at estimating the predictions of a wide range of fine-tuned BERT-family models and prompted GPT-family models, it is currently unclear when the Anchor Points technique can be expected to work. Specifically, when will the predictions of a target model follow the same linear trends as that of the set of source models? If the target

[1]Code is available at: https://drive.google.com/file/d/1FDHXGghi0Nls3O2NRBnK_Cilon50-Nty/view?usp=sharing

Preprint. Under review.

model or model family is known in advance, how should the source models be chosen to achieve the best estimation performance? We seek to answer these questions through analysis of predictive correlations across 53 fine-tuned and 20 in-context learned language models on four benchmark tasks.

We begin by reviewing the anchor points framework. We then investigate results suggesting that a target model whose predictions follow a sufficiently similar distribution to the source models will tend to share predictive correlations. The distribution difference can still be quite large, confirming that anchor point generalization does not rely on a target model making near identical predictions to one or more source models. We introduce the notion of a "prediction space" through which it becomes clear what target models are likely to be well-estimated using the anchor points of a given set of source models.

We then analyze the transfer performance of anchor points fit to a given family of models and used to estimate the predictions of models belonging to a different family. We define model "family" broadly as a set of models that share a distinct characteristic having a strong impact on behavior, e.g. models that all have a BERT-like architecture or models that are all instruction-tuned. (We note that families are not necessarily mutually exclusive). We find that anchor points do generalize across model families and gain some traction on our eponymous question. However, the best anchor point performance is always achieved when the source and target models belong to the same family. Finally, we recommend future directions for better understanding the phenomenon of predictive correlations in large models and how it can be leveraged to minimize model inference costs.

2 Related Works

2.1 Active Data Selection

The task of selecting a maximally representative or informative subset of data is well-explored in coreset (Guo et al. [2022]) and active data selection (MacKay, Ren et al. [2021], Kossen et al. [2021a] literature. Active data selection techniques often fall into one of two techniques: density-based sampling, where samples are chosen to sufficiently span the entire distribution of a data pool across some feature space (Guo et al. [2022]) and uncertainty-based sampling, where samples are chosen in areas where a model is most uncertain (Houlsby et al. [2011], Andreas Kirsch [2019]). The latter strategy has proven incredibly effective in active learning (AL), where the objective is to annotate the smallest number of training examples that allow a model to achieve sufficient generalization performance. Andreas Kirsch [2019] has achieved significant empirical success with strong theoretical underpinnings, while less rigorous uncertainty-based AL techniques have also shown strong empirical performance, a phenomenon elucidated in Farquhar et al. [2021].

As highlighted by Active Testing (Kossen et al. [2021a]), data selection for evaluation rather than training is a fundamentally different endeavor requiring distinct considerations of aleatoric and epistemic uncertainty as well as terms in the loss estimator. (Kossen et al. [2021a]) and (Kossen et al. [2021b]) achieve label-efficient model evaluation by using external “surrogate” models to approximate a model’s loss on unlabeled points, while others rely on topological summaries (Corneanu et al. [2020]) or performance on carefully-crafted synthetic data (Deng and Zheng [2021]). Importantly, these techniques aim to minimize the amount of *labels* required for model evaluation whereas Anchor Points aims to minimize the amount of *forward passes*.

2.2 Predictive Correlations at the Data Set Level

The phenomenon of consistent model predictive correlations at the data instance level is mirrored by prior work making analogous observations at the dataset level. Miller et al. [2021] highlighted the fascinating phenomenon: out-of-distribution (OOD) performance is strongly linearly correlated with in-distribution (ID) performance for a wide range of models and distribution shifts. These correlations hold for even for significant discrepancy in ID and OOD performance. This suggests that the choice of selecting a model with the best OOD performance reduces to selecting a model with the best ID performance. Agreement on the Line (Baek et al. [2022]) bring this convenience a step farther, showing that the agreement of any two models on an OOD task linearly correlates with their ID agreement. Thus, labels on OOD data are not required to estimate a model’s OOD performance. In general, these phenomena are not well-understood.

2.3 Predictive Correlations at the Data Point Level

To our knowledge, no works prior to Vivek et al. [2023] and this report draw attention to the inter-point correlations leveraged by Anchor Points. Zhong et al. [2021] emphasize that language model predictions are **noisy at the instance level**: training runs of the same model with different random seeds result in stochastic changes in instance-level predictions. This is true because inter-point correlations are not present across randomized instances of the same model but become obvious across the predictions of distinct models, a phenomenon closely related to Simpson's paradox. The authors do notice one correlative trend across models: improvement from BERT-Mini to BERT-Medium correlates with improvement from BERT-Medium to BERT-Large.

2.4 Imitation Learning to Minimize Inference Costs

Our approach is one among a wide variety of techniques that reduce the cost of model inference including quantization, knowledge distillation, pruning, sparsification, and efficient routing within Mixture-of-Experts. For brevity, we focus on the recent wave of large language model imitation learning, a kind of knowledge distillation. The rise of powerful proprietary models has led to many imitation efforts, where language models are trained on the outputs of larger and more expensive systems in an attempt to capture their behavior and performance (Wallace et al. [2020]). This is often performed in an analogous manner to Self-Instruct (Wang et al. [2023]), for example Alpaca (Taori et al. [2023]). While this approach seems broadly effective, recent work has questioned its effectiveness (Gudibande et al.). In our work, the predictive trends of weaker models can be used to estimate the predictions of much stronger proprietary models on a pre-specified dataset. Thus, our approach is analogous to imitation learning but works only in a transductive manner.

3 Problem Set-Up

3.1 Anchor Points Review

Anchor Points (Vivek et al. [2023]), is a technique that selects a subset of informative points from a labeled dataset such that a target model's predictions on the points can be used to estimate the model's predictions on many other points. Algorithm 1 and Algorithm 2 in Appendix A detail the process of fitting anchor points to the predictions of a set of source models S and then estimating the predictions of a target model T. In short, the distribution of points is represented by a space where distance $d(x_1, x_2) = 1 - PearsonCorrelation(\{(f_1(x_1), f_1(x_2))...(f_n(x_1), f_n(x_2))\})$ for source models $f_{1..n}$. K-Medoids is used in this space to select K points with good coverage of the distribution and then trend lines are fit between each test point and its closest anchor point. For estimating a target model's predictions on a given test point, the test point's trend line is evaluated on the corresponding target model's anchor point prediction. The authors found that anchor points achieve low error when fit and evaluated on fine-tuned BERT models as well as in-context learned GPT models. See Figures 7 and 8 in Appendix A for anchor point performance on GPT models.

3.2 Research Questions

1. How different can the predictions of model T be from the models in S without incurring large estimation error?
2. In practice, how can S be chosen to sufficiently estimate the predictions of a given T?

RQ1 asks whether anchor point generalization relies on source models making very similar predictions to target models, confirming that the technique does not have requirements that trivialize the problem.

RQ2 asks a generalized form of our eponymous question: "Can BERT Tell Me What GPT-3.5 will say?" It is of great practical interest whether less expensive models can be used to estimate the predictions of more expensive models.

Both of these questions work towards answering the fundamental question 'When will the predictions of a target model T fall on the trend lines that are fit to a set of source models S'? We gain insight but do not conclusively answer this question.

4 Methods

We perform a set of targeted experiments to answer the questions in Section 3.2:

4.1 How different can the predictions of S and T be while still obtaining low anchor point error?

We formalize the notion of the difference between the predictions of model S_1 and T by measuring the KL divergence between the models' predictions on each data point and then averaging over all data points in the validation set, i.e. $D(S_1, T) = \mathbb{E}_{x \in D_{val}}[KL(S_1(x), T(x))]$. The distance between the predictions of T and the predictions of the set of source models S can then be defined as $D(S, T) = \mathbb{E}_{S_n \in S}[D(S_n, T)]$.

4.1.1 Experiment 1: BERT-Family Generalization Analysis

1. Finetune 27 BERT-family models obtained from HuggingFace[2] each on four tasks from the GLUE benchmark Wang et al. [2019]: SST-2 (a binary sentiment classification task), CoLA, (a binary linguistic acceptability classification), MRPC (classification of whether a text passage paraphrases another test passage), and QNLI (classification of whether a paragraph contains the answer to a question). See the appendix for a full list of the models.
2. Obtain anchor point errors for each BERT model averaged over the validation set of each dataset in a leave-one-out manner. Specifically, anchor points are fit to the validation set predictions of all models except one and used to estimate the predictions of the held-out model.
3. Compute the mean KL divergence $D(B_n, B_m)$ for all model pairs in the dataset.
4. Represent the models in a continuous space using Multi-Dimensional Scaling where distance is mean KL divergence.
5. Look for trends between a model B_n's distance from other models and the Anchor Point estimation error of the B_n's predictions across the four datasets.

4.1.2 Experiment 2: Anchor Point Overfitting Analysis

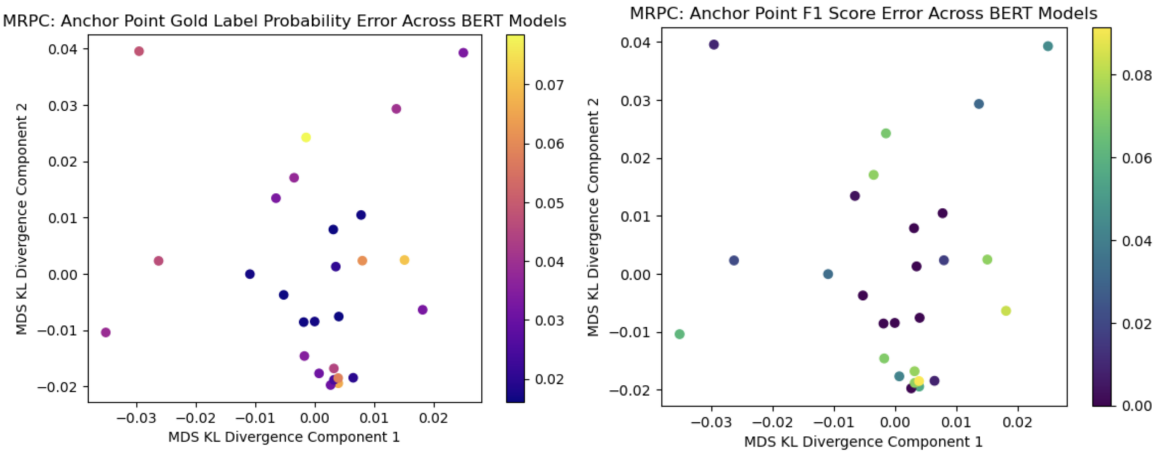
1. Finetune an additional 26 RoBERTa models (each with distinct pretraining data) on the same datasets. These serve as our set of source models S. See the appendix for a full list of the models.
2. Obtain anchor point errors for each BERT model from Experiment 1 using our RoBERTa source models S. Specifically, anchor points are fit to the validation set predictions of all models RoBERTa models and used to estimate the predictions of each BERT model.
3. Follow steps 3 and 4 of experiment 1 and look for trends between each BERT model B_n's distance from the RoBERTa models and the Anchor Point estimation error of the B_n's predictions across the four datasets.

4.2 How well can anchor points transfer across model families?

We define 5 model families each defined by a characteristic architecture, training, and/or size procedure. The full list of models can be found in Appendix A. Our model families are:

1. The BERT-family: 27 BERT-like models from Experiment 1
2. The RoBERTa-family: 26 RoBERTa models with distinct pretraining from Experiment 2
3. The (open-source) GPT-family: 10 GPT-like models from HuggingFace ranging from 100M - 1.3B parameters
4. The (open-source) InstructGPT family: 5 instruction fine-tuned GPT-like models ranging from 100M - 7B parameters
5. The OpenAI family: 3 GPT-3 models (text-ada-001, text-babbage-001, text-curie-001) and 2 GPT-3.5 models (text-davinci-002 and text-davinci-003)

[2]https://huggingface.co/models



(a) Colored by Error in Gold Label Probability Estimation.

(b) Colored by Error in Aggregate F1 Score Estimation.

Figure 1: Predictive Kullback Leibler Divergence Space of 27 BERT-family Models trained on MRPC, colored by the estimation error of 10 Anchor Points.

We then perform the following experiment:

4.2.1 Experiment 3: Anchor Point Transfer

1. Obtain predictions for the GPT, InstructGPT, and OpenAI families on the validation sets of CoLA and MRPC using three distinct prompts for each task. Prompts are obtained from PromptSource (Bach et al. [2022]).
2. For each pair of families, fit anchor points to the predictions of one family and obtain the estimation error of the predictions of all models belonging to the other family. For the GPT, InstructGPT, and OpenAI families, each model generates three sets of predictions for each task (due to the three different prompts).
3. Observe trends in the performance of transferred anchor points relative to non-transferred (leave-one-out) performance.
4. Compare the observed trends to notions of distance between the families, which can be defined using inter-cluster distance measurements in the mean KL-divergence space.

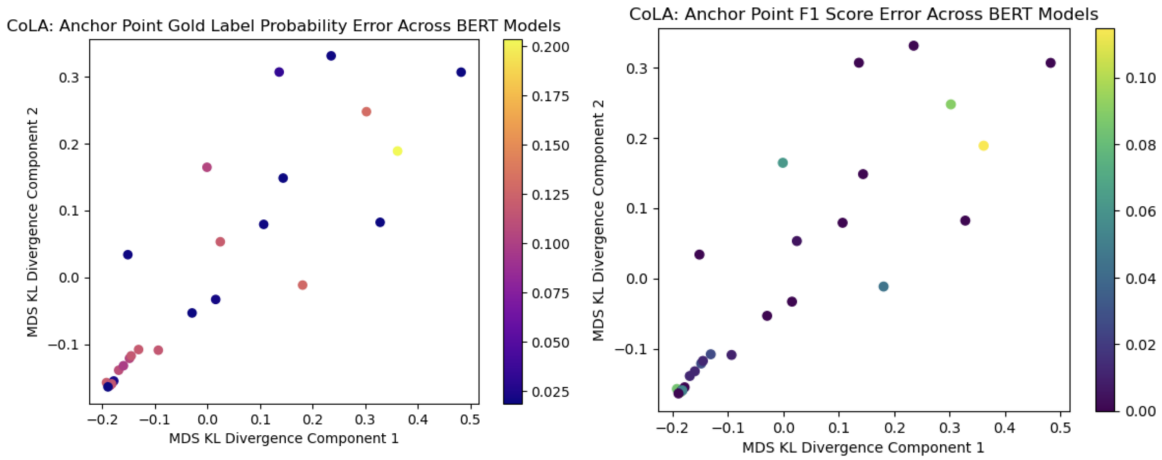
5 Empirical Results and Discussion

5.1 Experiment 1 (RQ1): Anchor points generalize across large predictive gaps.

We first analyze how well anchor points are able to generalize across predictive gaps, i.e. to models whose predictions deviate from the predictions of source models by a decent margin. Our method of visualizing models as points within a "prediction space" where distance is represented by mean KL divergence between different models' predictions proves useful. We report anchor point error in both estimated gold label (correct class) probability and aggregate macro-F1 score. We note that these errors are not always positively correlated. When models make unconfident predictions, even small gold label errors may cross the decision threshold, causing incorrect label flips and thus higher aggregate F1 estimation error.

We initially find that all 27 BERT models make very similar predictions for some datasets. For example, in Figure 1, we observe that all BERT-family models make predictions on the MRPC task that fall within a 0.04 bit radius sphere in the KL divergence space. Thus, estimating the predictions of a held-out BERT model is trivial as the mean prediction of the source models is already a strong approximation.

For other datasets, we find that anchor points generalize across large predictive gaps. For CoLA, anchor points achieve low error in estimated gold label probability and close to zero error in estimated



(a) Colored by Error in Gold Label Probability Estimation.

(b) Colored by Error in Aggregate F1 Score Estimation.

Figure 2: Predictive Kullback Leibler Divergence Space of 27 BERT-family Models trained on CoLA, colored by the estimation error of 10 Anchor Points.

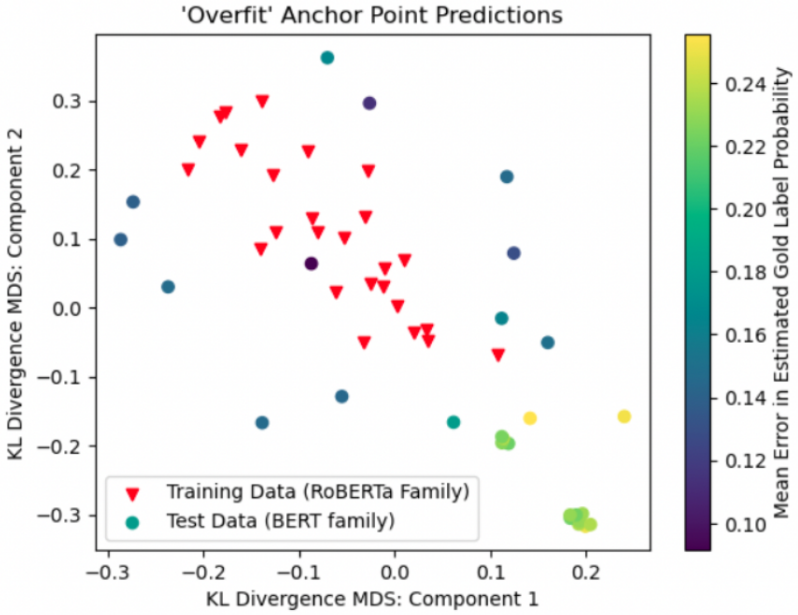


Figure 3: Predictive Kullback Leibler Divergence Space of 27 BERT-family and 26 RoBERTa family models trained on CoLA. Anchor points are fit to the RoBERTa source models and used to estimate the predictions of the BERT target models. Anchor point error tends to be higher on target models that fall farther from the source models.

macro-F1 score on most models. This is despite models spanning a much larger predictive region than in MRPC. For many models, the closest model is 0.1 bits away and others much farther. The low error obtained suggests that anchor points do not rely on target models making near identical predictions to source models. Rather, target models simply need to follow the same inter-example trends, a phenomenon that apparently occurs across models that make quite different predictions.

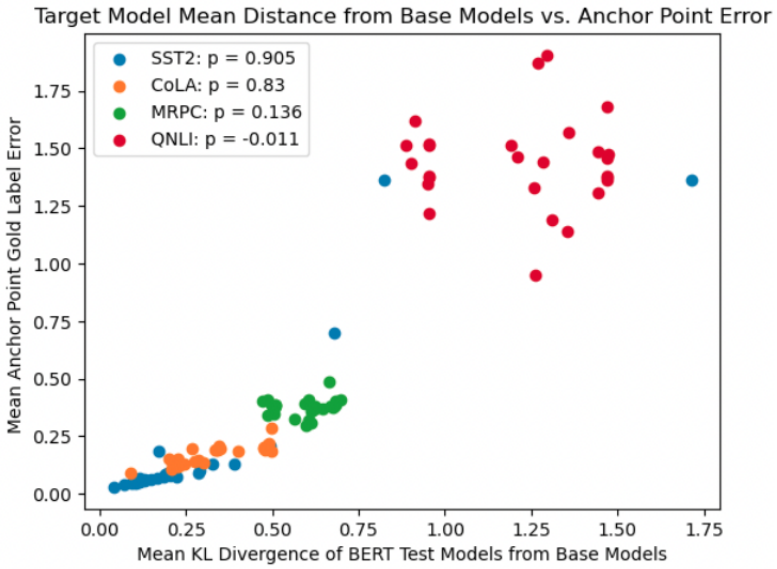


Figure 4: The KL divergence between target model predictions and the mean of source model predictions is plotted against anchor point estimation error across four datasets. A clear positive trend emerges, with greater noise as mean KL divergence increases. Intra-dataset Pearson correlations for each dataset are shown. Only SST-2 and CoLA achieve statistically significant linear trends.

5.2 Experiment 2 (RQ1): Anchor point generalization deteriorates as predictive gaps between source and target models grow.

We next investigate the limits of anchor point generalization. We "overfit" anchor points by fitting them to the predictions of 26 RoBERTa-family models and then evaluate their estimates on BERT-family model predictions. We note that RoBERTa-family models are technically a subset of BERT-family models. This fact is mirrored by our predictive space shown in Figure 3, where RoBERTa models fall within a sub-region of the broader BERT predictive space. We observe that anchor point gold label errors are substantially higher than in Figure 2. Moreover, anchor point generalization tends to deteriorate for target models that are farther from the source models.

To further confirm the claim that anchor point generalization deteriorates across larger predictive gaps, we plot the KL Divergence between a given target model and source models (averaged over all source models and data points) against the anchor point gold label error on the target model. A clear positive trend appears, where anchor point error appears to be lower bounded by an exponential.

We note that anchor point generalization is not a function of only mean KL divergence between source models and a given target model. Some BERT models in Figure 3 have a smaller predictive gap to source models than that of models in Figure 2, yet prove more difficult to estimate. The clear trend in Figure 4 appears only when conditioned on a given choice of source and target models.

5.3 Experiment 3 (RQ2): Anchor points generalize across model families to varying degrees.

In our final experiment, we assess the performance of transferring anchor points between model families. Our results on CoLA appear in Figures 5 and 6. Overall, we observe that anchor points fit to one model family often do have predictive power on another model family, albeit weaker than that of anchor points fit to the other family. We also observe that transfer performances between families change between tasks: transfer performances on MRPC are quite distinct (see Figures 9 and 10.

5.3.1 How should source models S be chosen to estimate a target model T?

It is clear that the most reliable choice of source models S for a given model T is to choose S from the same family as T. While some family pairs attain small transfer F1 errors, anchor points fit to the same model family achieve lower error at the instance level by a large margin. However, in

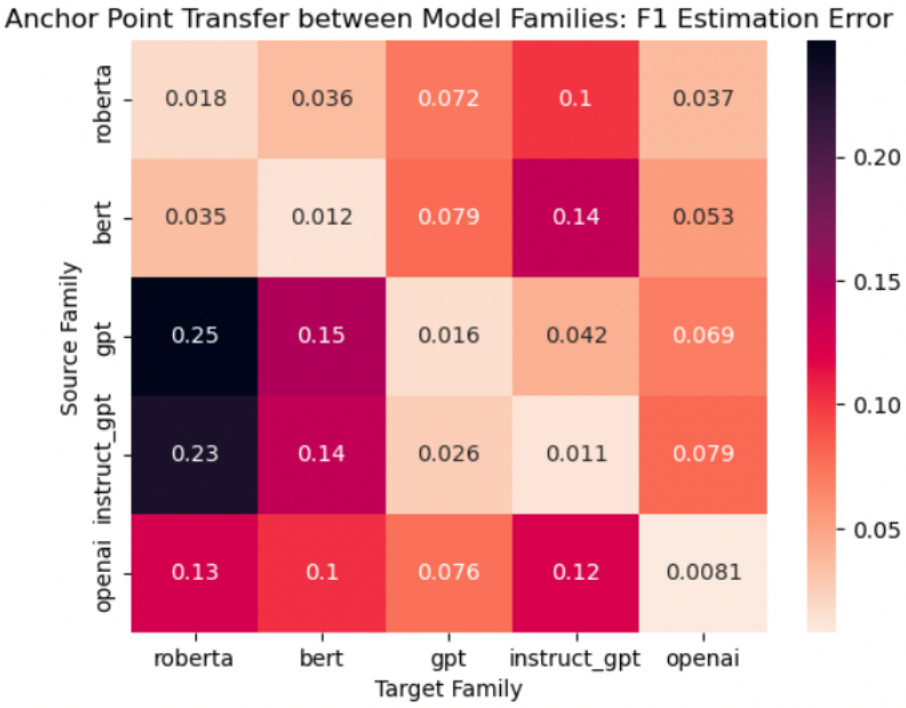


Figure 5: Anchor point transfer performance across model families. Results are aggregate F1 estimation errors averaged over three prompts on CoLA. For scores along the diagonal, leave-one-out error is computed as in Experiment 1.

situations where only aggregate level performance estimation is needed, taking advantage of anchor point transfer may prove effective.

5.3.2 Can BERT tell me what GPT-3.5 will say?

To some extent, yes. The last two columns of the transfer grid in Figure 11 show some success of BERT-family anchor points estimating the predictions of text-davinci-002 and text-davinci-003, but this varies with the prompt and dataset. Generally, it seems that transferred anchor points get gold label estimation errors in the range of 0.15 - 0.3. When target models make sufficiently confident predictions, this instance-level error is low enough to estimate the correct class consistently. Thus, transferring anchor points from cheaper, open-source models to more expensive, perhaps proprietary models may be practical.

6 Conclusions

We empirically demonstrate that language models tend to share inter-example predictive correlations when they have sufficiently similar predictive distributions. The anchor points technique can generalize across fairly large predictive gaps between source and target models, but generalization deteriorates as the gap grows. In practice, we find that the best source models for estimating the predictions of a given target model are models that belong to the same family as the target model, i.e. that have a similar architecture and training set-up.

We recommend a few important directions to explore in future work. On the practical side, we hope to investigate anchor point behavior on regression tasks as well as adversarial tasks, both of which may exhibit different characteristics. We also think that building sets of source models specifically to estimate the predictions of popular proprietary models may be a promising endeavor. On the theoretical side, we hope that the research community works toward better theoretical understanding of the phenomenon of inter-example predictive correlations.

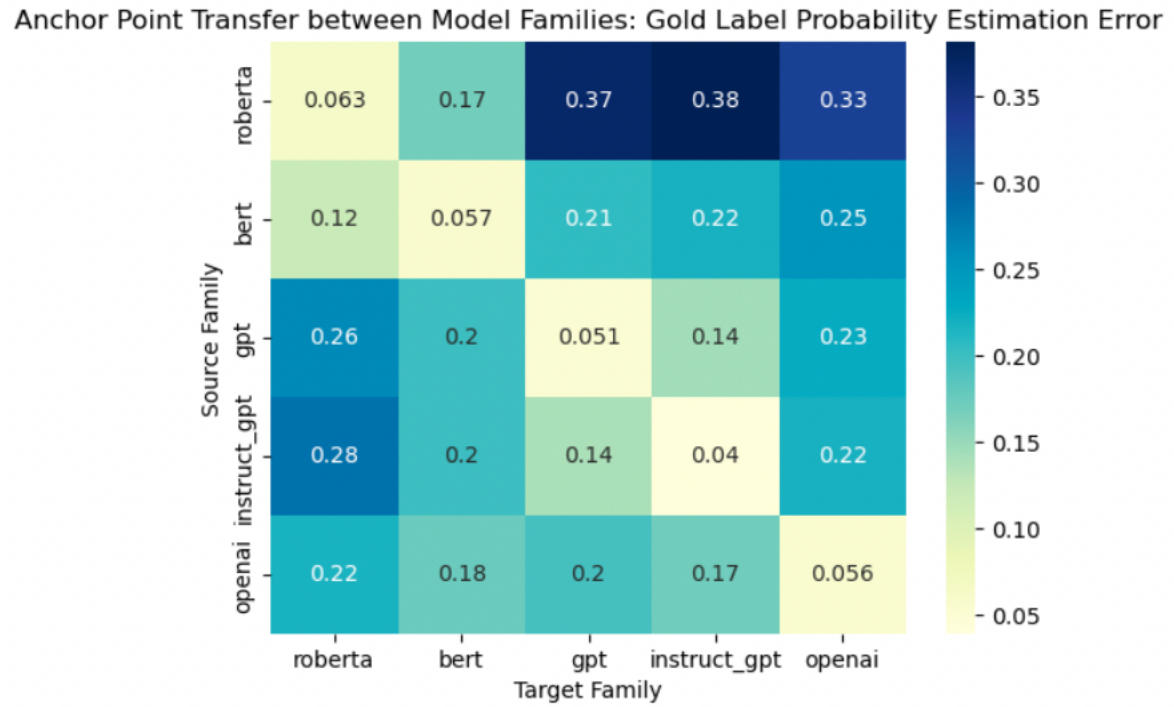


Figure 6: Anchor point transfer performance across model families. Results are gold label probability estimation errors averaged over three prompts on CoLA. For scores along the diagonal, leave-one-out error is computed as in Experiment 1.

To accelerate these endeavors, we recommend that members of the research community begin releasing model predictions at the instance level on widely-used datasets. Understanding the phenomena underlying correlative behavior across language models will likely lead to improved techniques for reducing inference costs and insights for better developing machine learning systems.

7 Contributions and Commentary on a Related Project

This work was advised by Douwe Kiela, Diyi Yang, and Kawin Ethayarajh. All implementation effort and writing is my own.

I used other Anchor Points-related work for my final project submission to CS 224U this quarter. However, this project and the 224U project are distinct with disjoint sets of experiments and results. The 224U project focused on building the anchor points technique and comparing its performance to baselines. The only content in this paper that is shared with the 224U paper (besides the central anchor point technique) is placed in the Appendix: Algorithm 1+2 and Figures 7+8.

8 Appendix A

8.1 Models

The BERT, RoBERTa, GPT, and InstructGPT models can be found on https://huggingface.co/models.

8.1.1 BERT Family Models

1. 'Bio_ClinicalBERT'
2. 'albert-base-v2'
3. 'bart-base'
4. 'bart-large'

Algorithm 1 Anchor Points Fit

$k \leftarrow$ number of anchor points
$P \leftarrow NxD$ array of N source models' normalized gold label predictions on D points
$C \leftarrow$ corrcoef(P) ▷ $D \times D$ correlation matrix
$AP \leftarrow$ K-MEDOIDS(1 - C, K) ▷ length K array of anchor point indices
$N \leftarrow$ argmax(C[:, AP], axis = 1) ▷ length D array of nearest anchor point to each point in D
$T \leftarrow \{0...D-1\} \setminus AP$ ▷ length D - K array of test point indices
$M \leftarrow$ empty $K \times (D-K)$ array ▷ Parameter matrix of slopes
$B \leftarrow$ empty $K \times (D-K)$ array ▷ Parameter matrix of biases
for i **do**, anchor $\in$ enumerated AP
  for j **do**, test $\in$ enumerated T
    m, b $\leftarrow$ LinearRegression(P[:,anchor], P[:,test]) ▷ Slope and bias of trend line
    M[i, j] $\leftarrow$ m
    B[i, j] $\leftarrow$ b
  end for
end for
  return AP, N, T, M, B

Algorithm 2 Anchor Points Predict

Require: AP, N, T, M, B ▷ Returned by Anchor Points Fit
P $\leftarrow$ length K array of a target model's gold label predictions on the K anchor points
Y $\leftarrow$ empty length D array ▷ To store estimated target model predictions
Y[AP] $\leftarrow$ P
for i, anchor $\in$ enumerated AP **do**
  for j, test $\in$ enumerated T **do**
    Y[i,j] $\leftarrow$ M[i, j] * P[i] + B[i,j] ▷ Estimated Prediction
  end for
end for
Y $\leftarrow$ Y[N, [0...length(Y) - 1]]
▷ For each test point, keep only the prediction of its nearest anchor point
  return Y

5. 'bert-base-cased'
6. 'bert-base-multilingual-cased'
7. 'bert-base-uncased'
8. 'bert-large-cased'
9. 'bert-large-uncased'
10. 'bert-mini'
11. 'bert-tiny'
12. 'biobert-v1.1'
13. 'deberta-base'
14. 'deberta-large'
15. 'deberta-v3-base'
16. 'deberta-v3-xsmall'
17. 'distilbert-base-cased'
18. 'distilbert-base-uncased'
19. 'electra-base-discriminator'
20. 'legal-bert-small-uncased'
21. 'roberta-base'
22. 'scibert_scivocab_uncased'

23. 'sentence_bert'
24. 'sentiment-roberta-large-english'
25. 'twitter-roberta-base'
26. 'xlm-roberta-base'
27. 'xlm-roberta-large'

8.1.2 RoBERTa Family Models

1. 'biomed_roberta_base',
2. 'cs_roberta_base',
3. 'dsp_roberta_base_dapt_biomed_tap_chemprot_4169',
4. 'dsp_roberta_base_dapt_biomed_tapt_rct_180K',
5. 'dsp_roberta_base_dapt_biomed_tapt_rct_500',
6. 'dsp_roberta_base_dapt_cs_tapt_citation_intent_1688',
7. 'dsp_roberta_base_dapt_cs_tapt_sciie_3219',
8. 'dsp_roberta_base_dapt_news_tapt_ag_115K',
9. 'dsp_roberta_base_dapt_news_tapt_hyperpartisan_news_5015',
10. 'dsp_roberta_base_dapt_news_tapt_hyperpartisan_news_515',
11. 'dsp_roberta_base_dapt_reviews_tapt_amazon_helpfulness_115K',
12. 'dsp_roberta_base_dapt_reviews_tapt_imdb_20000',
13. 'dsp_roberta_base_dapt_reviews_tapt_imdb_70000',
14. 'dsp_roberta_base_tapt_ag_115K',
15. 'dsp_roberta_base_tapt_amazon_helpfulness_115K',
16. 'dsp_roberta_base_tapt_chemprot_4169',
17. 'dsp_roberta_base_tapt_citation_intent_1688',
18. 'dsp_roberta_base_tapt_hyperpartisan_news_5015',
19. 'dsp_roberta_base_tapt_hyperpartisan_news_515',
20. 'dsp_roberta_base_tapt_imdb_20000',
21. 'dsp_roberta_base_tapt_imdb_70000',
22. 'dsp_roberta_base_tapt_rct_180K',
23. 'dsp_roberta_base_tapt_rct_500',
24. 'dsp_roberta_base_tapt_sciie_3219',
25. 'news_roberta_base',
26. 'reviews_roberta_base'

8.1.3 (Open Source) GPT Family Models

1. 'Cerebras-GPT-1.3B'
2. 'Cerebras-GPT-111M'
3. 'Cerebras-GPT-256M'
4. 'bloom-1b7'
5. 'gpt-neo-1.3B'
6. 'gpt-neo-125m'
7. 'gpt2-large'
8. 'gpt2-medium'
9. 'gpt2'
10. 'openai-gpt'

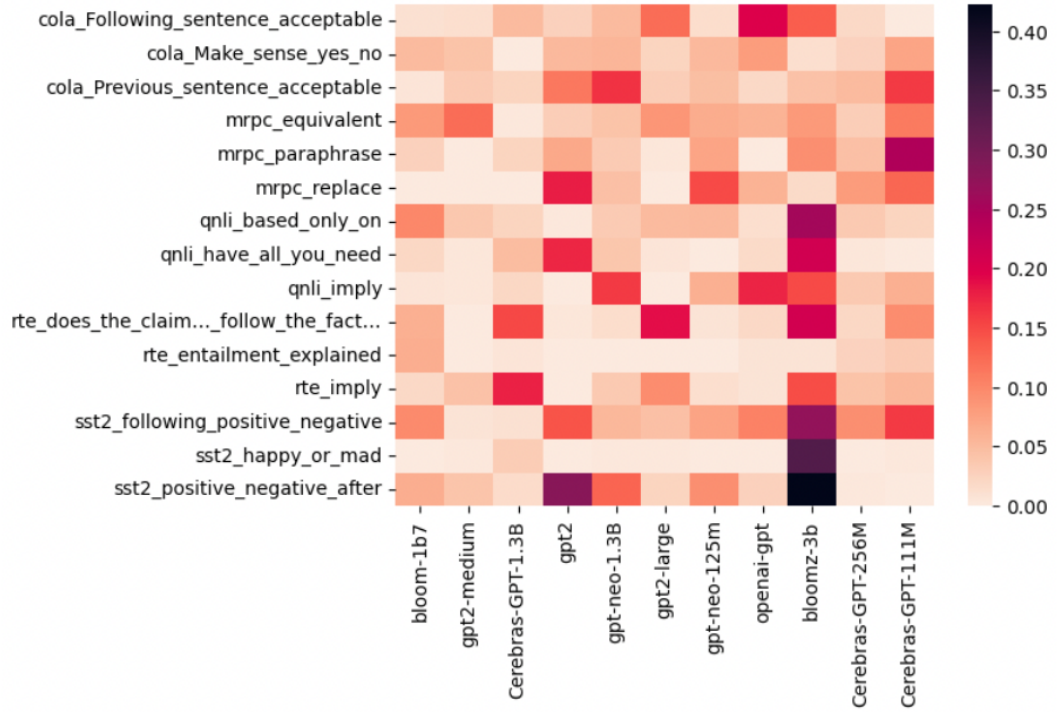


Figure 7: Error in aggregate F1 score estimation of GPT-family anchor points computed in a leave-one-out style. The rows refer to different tasks and PromptSource prompts. Note that bloomz-3b is the only instruction-tuned model, a likely explanation for poor anchor point generalization to its predictions.

8.1.4 (Open Source) InstructGPT Family Models

1. 'RedPajama-INCITE-Instruct-7B-v0.1'
2. 'falcon-7b-instruct'
3. 'mpt-7b-instruct'
4. 'mt0-xl'
5. 'bloomz-3b'

8.1.5 OpenAI Family Models

1. 'text-ada-001',
2. 'text-babbage-001',
3. 'text-curie-001',
4. 'text-davinci-002',
5. 'text-davinci-003'

References

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Exploiting instance-level correlations for efficient model benchmarking and prompt selection. 2023.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

Ryan Smith, Jason Fries, Braden Hancock, and Stephen Bach. Language models in the loop: Incorporating prompting into weak supervision. 2022.

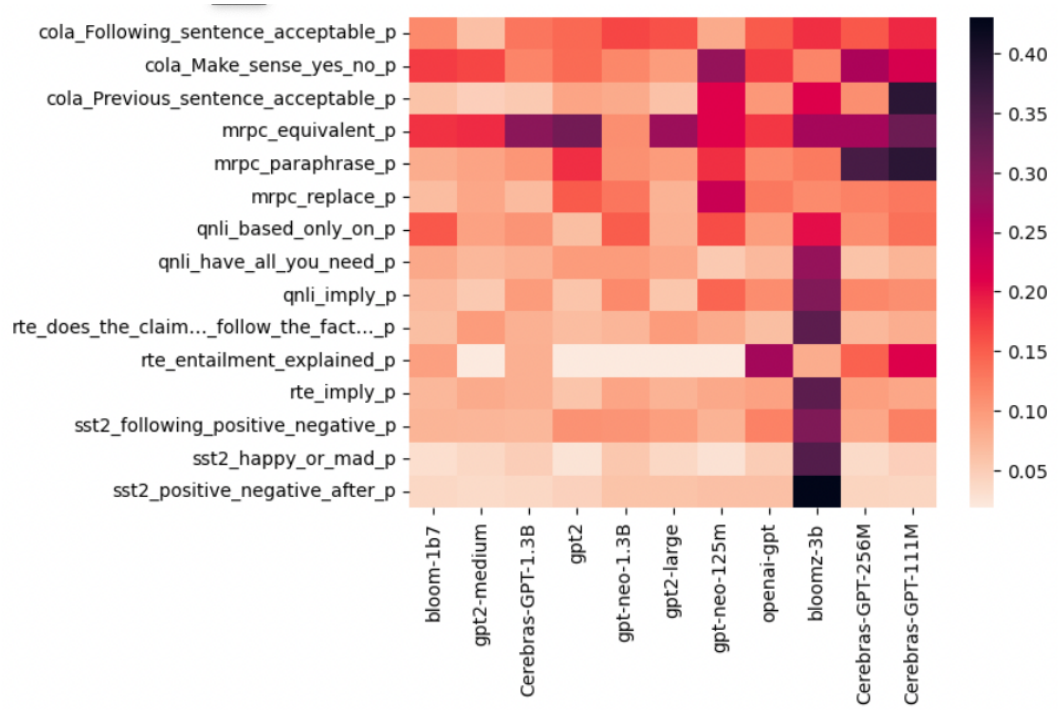


Figure 8: Error in gold label probability estimation of GPT-family anchor points computed in a leave-one-out style. The rows refer to different tasks and PromptSource prompts. Note that bloomz-3b is the only instruction-tuned model, a likely explanation for poor anchor point generalization to its predictions.

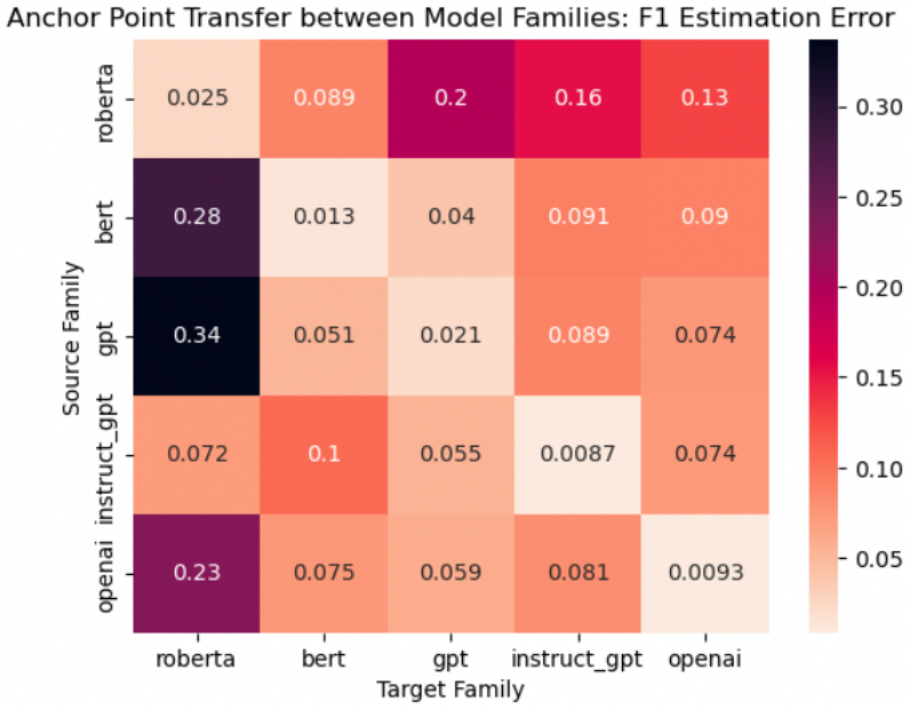


Figure 9: Anchor point transfer performance across model families. Results are aggregate F1 estimation errors averaged over three prompts on MRPC. For scores along the diagonal, leave-one-out error is computed as in Experiment 1.

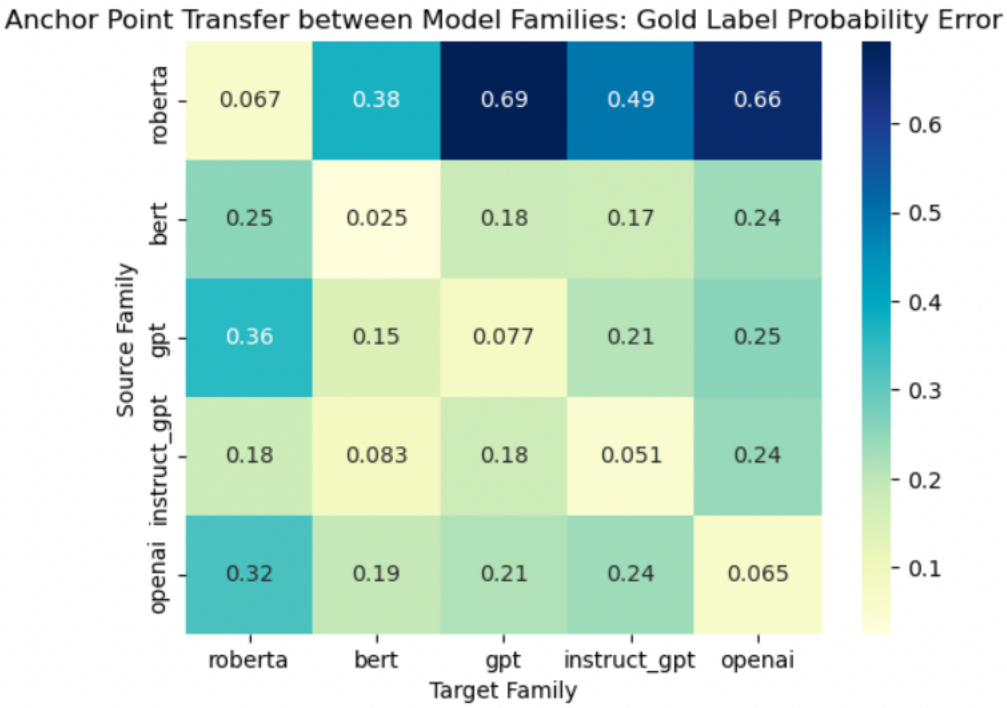


Figure 10: Anchor point transfer performance across model families. Results are gold label probability estimation errors averaged over three prompts on MRPC. For scores along the diagonal, leave-one-out error is computed as in Experiment 1.

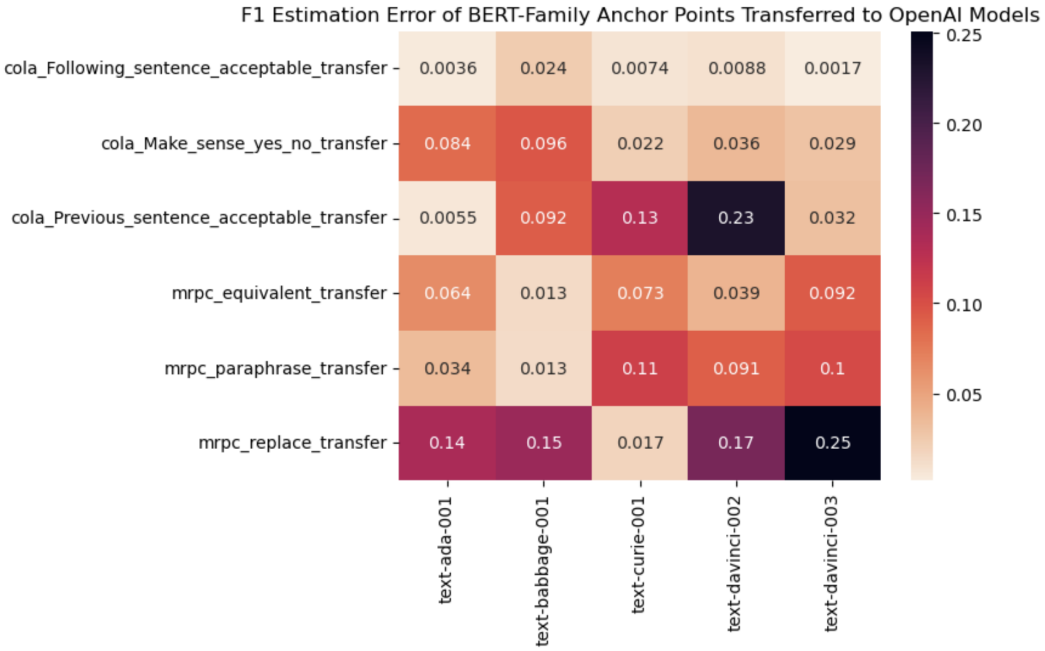


Figure 11: Error in Aggregate F1 Score of BERT-family Anchor Points transferred to OpenAI Models. The rows refer to different tasks and PromptSource prompts.

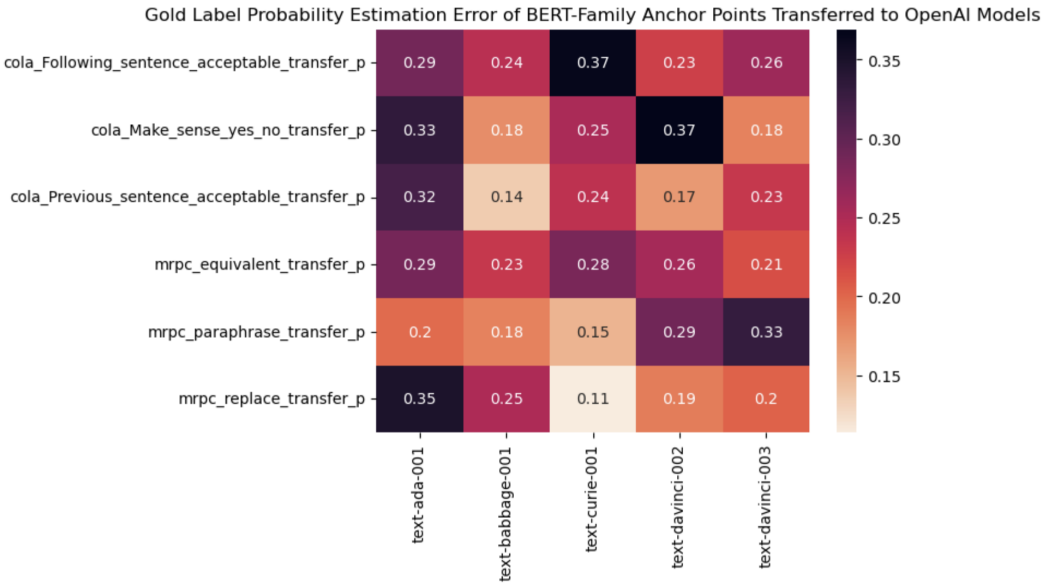


Figure 12: Error in Gold Label Probability of BERT-family Anchor Points transferred to OpenAI Models. The rows refer to different tasks and PromptSource prompts.

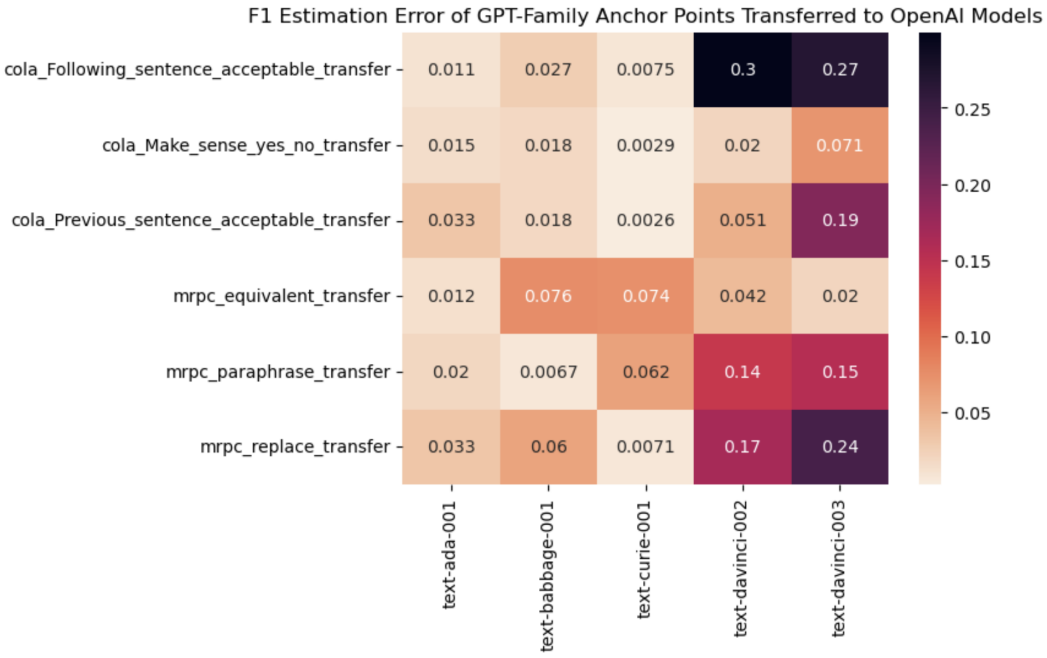


Figure 13: Error in Aggregate F1 Score of GPT-Family Anchor Points transferred to OpenAI Models. The rows refer to different tasks and PromptSource prompts.

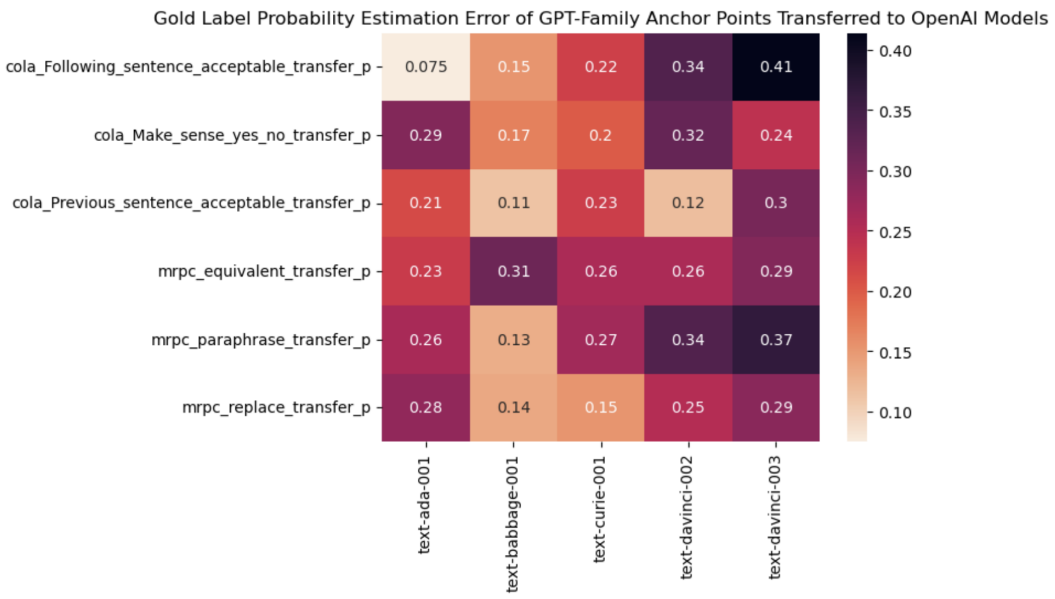


Figure 14: Error in Gold Label Probability of GPT-family Anchor Points transferred to OpenAI Models. The rows refer to different tasks and PromptSource prompts.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. 2022. doi: https://doi.org/10.48550/arXiv.2204.08499.

David J. C. MacKay. Information-based objective functions for active data selection.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. 2021.

Jannik Kossen, Sebastian Farquhar, Yarin Gala, and Tom Rainforth. Active testing: Sample-efficient model evaluation. 2021a.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. 2011.

Yarin Gal Andreas Kirsch, Joost van Amersfoort. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. 2019.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. 2021.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active surrogate estimators: An active learning approach to label-efficient model evaluation. 2021b.

Ciprian Corneanu, Meysam Madadi, Sergio Escalera, and Aleix Martinez. Computing the testing error without a testing set. 2020.

Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? 2021.

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. 2021.

Christina Baek, Yiding Jiang, Aditi Raghunathan, and Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. 2022.

Ruiqi Zhong, Dhruba Ghosh, Dan Klein, and Jacob Steinhardt. Are larger pretrained language models uniformly better? comparing performance at the instance level. 2021.

Eric Wallace, Mitchell Stern, and Dawn Song. Imitation attacks and defenses for black-box machine translation systems. *EMNLP*, 2020.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. 2023.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms. *arxiv*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. 2019.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, et al. Promptsource: An integrated development environment and repository for natural language prompts. 2022.