Do we still need BERT in the age of GPT? Comparing the benefits of domain-adaptation and in-context-learning approaches to using LLMs for Political Science Research

Mitchell Bosley*[1], Musashi Jacobs-Harukawa†[2], Hauke Licht‡[3], and Alexander Hoyle§[4]

[1]University of Michigan
[2]Princeton University
[3]University of Cologne
[4]University of Maryland

April 14, 2023

Abstract

With the rapid development of large language models (LLMs), we claim that researchers using LLMs must make three critical decisions: model selection, domain-adaptation strategies, and prompt design. To help provide guidance on these choices, we establish a set of benchmarks for a wide range of natural language processing (NLP) tasks pursued by political science tasks. We use this benchmark to compare two common approaches to the classification of political text: domain-adapting smaller LLMs such as BERT to one's own data with varying levels of unsupervised pre-training and supervised fine-tuning, and querying larger LLMs such as GPT-3 without additional training. Preliminary results indicate that when labeled data is available, the fine-tuning focused approach remains the superior technique for text classification.

1 Introduction

Traditionally, supervised and unsupervised machine learning methods have been used in the realm of natural language processing (NLP) for quantifying concepts in political text. Each approach has its benefits and drawbacks: supervised methods yield easily interpretable results but necessitate costly labeling, while unsupervised approaches, such as Latent Dirichlet Allocation (LDA Blei et al., 2003), are cost-effective but challenging to interpret. With the advent of advancements in the field of machine learning and artificial intelligence such as neural networks, word embeddings, transformers, and large language models (LLMs), transfer learning, which combines both unsupervised and supervised approaches, has become the

*PhD Candidate, Department of Political Science. mcbosley@umich.edu
†Postdoctoral Researcher, Data-Driven Social Science Initiative. mjacobsharukawa@princeton.edu
‡Postdoctoral Researcher, Center for Comparative Politics. hauke.licht@wiso.uni-koeln.de
§PhD Candidate, Department of Computer Science. hoyle@umd.edu

standard in applied machine learning (Pan & Yang, 2010; Zhuang et al., 2020). In the transfer learning approach, LLMs such as BERT[1] (Devlin et al., 2019) are first *pre-trained* on a large general text corpus on an unsupervised task such as masked-word prediction, and then *fine-tuned* using a supervised approach on main-specific natural language processing (NLP) tasks such classification. While this approach has proven successful in achieving state-of-the-art performance across a wide array of NLP tasks (Raffel et al., 2020), the boundary at which additional unsupervised pretraining vs. finetuning with one's own data remains unclear (Gururangan et al., 2020).

As the data and computational resources used to train LLMs has grown, so too has their language processing proficiency. Generative Pre-Trained Transformer (GPT) models have been trained to generate text in response to any arbitrary prompt by iteratively predicting the word that is most likely to come next (Radford et al., 2019). GPT models like InstructGPT (Ouyang et al., 2022) have been shown to accomplish a variety of NLP tasks with only a few examples in its prompt, a technique referred to as *in-context learning* (ICL, Brown et al., 2020a). In-context learning, or prompt-engineering, involves manipulating the information provided to an LLM via its prompt to enhance the model's performance on a specific task. However, ICL remains more art than science, and determining the optimal information to provide is an active research area with conflicting results (Min et al., 2022).

There are also financial and ethical concerns associated with using GPT-class models. While these larger models can also be fine-tuned for domain-specific tasks, the computational burden is significantly higher than for smaller models like BERT. Moreover, since the computational cost of producing outputs from high-capability GPT models like GPT-3 and GPT-4 is high, researchers are forced to use proprietary and closed-source models from tech companies like OpenAI, raising concerns about reproducibility and cost.

Currently, then, it is an open question as to whether fine-tuning a relatively small model such as BERT for a target task is a better approach than using in-context learning to produce predictions from a large general-purpose language model like GPT-4 whose size prohibits fine-tuning.

Our central claim is that any researcher using LLMs for text analysis must make three discrete choices: (1) the *model* to use, (2) the *training strategy* to employ for domain-adaptation, and (3) and the approach to *in-context-learning*. Our goal is to provide guidance to researchers navigating these choices. Towards this end, we have begun the process of establishing a set of benchmarks for various political science NLP tasks. As an initial investigation, we use these benchmarks to compare the benefits of two approaches: domain-adapting smaller LLMs (such as BERT) to one's own data, and using in-context learning with larger LLMs like GPT-3 or GPT-4 without additional training. Our preliminary results indicate that when labeled data is available for fine-tuning, it provides substantially better performance for classification tasks than zero-shot or in-context learning approaches using GPT-3. We also show that when labeled data is plentiful, additional pre-training does not meaningfully improve performance.

This paper is organized as follows. In Section 2, we offer a brief review of the Transformers architecture and text classification using LLMs. In Section 3 we describe our process of benchmark construction (including dataset selection and description), model architectures, training strategies, evaluation metrics, and experimental design. In Section 4, we present our preliminary results, comparing the performance of different approaches for classification across using our benchmarks. Lastly, in Section 5, we discuss our findings and provide

[1] Bidirectional Encoder Representations from Transformers

recommendations for practitioners.

2 Large Language Models for Text Classification

2.1 The Transformer Architecture

The development of large language models (LLMs) and the Transformer architecture has been driven by a series of innovations in machine learning, such as word embeddings, neural networks, and the Transformer architecture itself.

Word embeddings are dense vector representations of words that capture their semantic meaning in a continuous space. These representations are learned from large corpora of text data and have proven to be effective in various natural language processing tasks. Early word embedding techniques, such as Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), enabled researchers to capture semantic relationships between words, providing an alternative to traditional one-hot encodings or bag-of-words representations.

Neural networks provided the foundation for increasingly complex models capable of learning from and generating text data (Elman, 1990; Hochreiter & Schmidhuber, 1997; Kim, 2014). These advancements paved the way for the groundbreaking Transformer architecture, introduced by Vaswani et al. (2017), which combined the strengths of previous neural network models to create a powerful, scalable, and highly effective neural network for various NLP tasks, including text classification.

The Transformer architecture is composed of two main components: an encoder, which processes the input text by breaking it down into meaningful representations, and a decoder, which generates the output text based on these representations. The Transformer is based on a combination of self-attention mechanisms, which help the model focus on relevant parts of the text, and feed-forward layers, which process the information in a single direction without looping back. These components are organized in a stacked arrangement, forming multiple layers that enable the model to learn complex relationships between words. To account for the order of words in a sequence, the Transformer incorporates positional encoding, which adds information about the position of words, allowing the model to recognize and learn patterns based on word order. This design allows the Transformer to efficiently identify relationships between words in a sequence, regardless of their distance from one another. The larger scale is crucial, as a key lesson from decades of machine learning research is *the more, the better* (Halevy et al., 2009).

2.2 Fine-tuning with large pretrained language models

BERT (Bidirectional Encoder Representations from Transformers) is a powerful language model based on the Transformer architecture's encoder, designed for a wide range of natural language processing tasks, including text classification (Devlin et al., 2019).

BERT is pre-trained with a *masked language modeling* objective, wherein a certain percentage of words in a sentence are masked, and the model is trained to predict the masked words based on their surrounding context. This pre-training step allows BERT to learn bidirectional contextual representations, as it must understand the context of words from both the left and right.

After pre-training, BERT models can be *fine-tuned*: the general language knowledge acquired pre-training supports the subsequent learning of specific tasks, such as text classification. This approach has proven effective across various NLP tasks, as the pre-trained knowledge can be transferred and fine-tuned to the target domain.

Several BERT-like models have emerged, offering variations and improvements over the original BERT architecture, with notable examples such as RoBERTa and DistilBERT.[2] Although there are hundreds of BERT-like models that have surpassed the originals' performance on a variety of tasks, the basic pretrain-and-finetune paradigm remains dominant.

2.3 In-context learning with GPT models

The Generative Pre-trained Transformer (GPT, Radford et al., 2018) family of LLMs, such as GPT-3 (Brown et al., 2020a) and its successor GPT-4 (OpenAI, 2023), are state-of-the-art autoregressive language models that perform exceptionally well across a wide range of natural language processing tasks, including text classification.

There are three critical differences between BERT and GPT families of models: their training objectives, the computational scale at which they are trained, and their prompt size. BERT class models are mostly trained using masked language modeling, where it predicts missing words in a sentence given their context.[3] GPT-3, on the other hand, is trained exclusively to predict the next word in a sequence, also known as causal language modeling.

In terms of scale, GPT-3 has 175 billion parameters, which is orders of magnitude larger than BERT's 340 million parameters. The larger scale allows GPT-3 to capture more information from the vast amounts of data used during pretraining, which is estimated to be hundreds of gigabytes of text (Brown et al., 2020b). The substantial scale of the data and computation used to train GPT-3 lead to "emergent" generalization behaviors. This increased capacity allows GPT models to better utilize the available context window when processing text.

Larger context windows in GPT models enable more effective in-context learning. While the context window size is primarily determined by the model architecture and configuration, the model scale and training objective influence the model's capacity to effectively utilize the context window. GPT models' larger scale and unidirectional context learning from causal language modeling allow them to process longer sequences and better handle long-range dependencies.

The combination of the massive scale of the data and computation used to train the model with the causal model objective and increased context windows has made GPT models more adept at *generating* text given a prompt than BERT models, and has lead to "emergent" generalization behaviors. GPT-3 has demonstrated remarkable zero and few-shot learning abilities, meaning it can perform tasks without explicit fine-tuning or with only a few examples provided as context, and early indications are that GPT-4 is even more adept (Bubeck et al., 2023). To perform few-shot learning with GPT models, users provide examples of the desired input-output pairs as part of the prompt, which helps the model learn

[2] RoBERTa (Robustly optimized BERT approach) proposed by Liu et al. (2019), improves upon BERT by modifying the training procedure and using larger mini-batches, longer training time, and removing a next sentence prediction task, leading to improved performance across a range of tasks. DistilBERT, introduced by Sanh et al. (2019), a smaller, faster, and more efficient version of BERT, achieved through knowledge distillation. DistilBERT retains approximately 95% of BERT's performance while reducing the model size by 40% and requiring 60% fewer computations.

[3] The original BERT was trained on both masked-language and next-sentence prediction tasks, but subsequent BERT-like models such as RoBERTa showed that dropping the next-sentence prediction task in favor of more masked-language modeling led to better results.

the format of the task and generate appropriate responses. The prompt can also include any additional context or instructions necessary for the model to complete the task. Table 1 shows an example of in a political text classification task using in-context learning, where a user provides GPT-3 several labeled examples, specifying the classification categories and a brief instruction on how to classify the text. The model then generates a classification label for the given input text, based on the examples and instructions provided.

Input Text	**Output Label**
Example 1: Government should provide universal health care to all citizens.	Liberal
Example 2: Lower taxes stimulate economic growth.	Conservative
Example 3: The weather today is sunny with a high of 75 degrees Fahrenheit.	Neutral
...	...
New Input: Regulations on businesses should be reduced to encourage entrepreneurship.	[To be predicted by the model]

Table 1: Example of in-context learning with a GPT model for a political text classification task. The model is provided with labeled examples and is expected to classify the new input text based on the context from these examples.

2.4 Closed source concerns

An important additional consideration when choosing between the pretrain-finetune paradigm with BERT models and the pretrain-prompt engineering approach with GPT models is the closed-source nature of the most capable GPT models. Unlike the majority of BERT models, which are open-source and easily accessible, the largest GPT models are not publicly available due to their immense size, computational requirements, and potential misuse. As a result, researchers and practitioners must rely on APIs provided by the organizations maintaining these models (such as OpenAI), which can lead to limitations in terms of access, customization, and reproducibility.

The closed-source nature of the most capable GPT models raises concerns for reproducible research. Since these models are not readily available for experimentation, it becomes challenging for researchers to reproduce and verify the results of studies that use these models. Moreover, the lack of access to the underlying model and training data makes it difficult for researchers to investigate potential biases, limitations, and opportunities for improvement in these models.

BERT models, in contrast, are typically open-source, allowing researchers to access, adapt, and fine-tune them as needed, which facilitates reproducible research.

2.5 The large and the very large

Despite the recent attention paid to GPT-type models and their surprising abilities, the computer science and machine learning literature has not found prompting large GPT-class models to be a definitively superior alternative to the pretrain-and-finetune paradigm (Liu et al., 2022). Although some findings suggest that prompting is more robust to spurious correlations in training data, it can also fail to learn from the examples provided in the prompt, performing identically when presented with randomized labels (Si et al., 2023). Indeed, there are many tasks in which large language models fail to surpass weak baselines(OpenSamizdat, 2021).

It is therefore not a given that we should abandon domain-adaptation with smaller models like BERT in favor of prompt-engineered GPT models. Rather, we should be careful to evaluate which approach our task is best suited to. In the next section, we outline our approach to investigating whether additional pretraining on domain-specific data can help smaller masked-language-models (such as BERT) outperform their larger unadapted counterparts (such as GPT) in the context of the three-part choice we outline. By exploring the trade-offs between model choice, domain adaptation, and prompt design, we aim to provide a clearer understanding of when each paradigm might be most suitable for a given task or domain.

3 Methodology

This section describes the research design we use to evaluate the performance of the two paradigms of text classification: domain adaptation with BERT-like models, involving a combination of pretraining with masked language modeling and finetuning with supervised classification, and in-context learning with GPT models. In Section 3.1 we discuss our ongoing efforts at constructing a benchmark evaluating methods of analyzing political text, including large unlabeled datasets to use for unsupervised pre-training of LLMs, and labeled datasets to use for supervised fine-tuning. In Section 3.2 we describe the models we use for our experiments, including the BERT models we fine-tune and the GPT models we prompt. Then, in Section 3.3 we describe the experiments we conduct to evaluate the performance of the two paradigms.

3.1 Dataset selection and description

3.1.1 Pre-training data

For model pre-training, we have compiled a novel dataset of English-language parliamentary speech records from the lower houses of Australia (*House of Representatives*, 1998-2012), Canada (*House of Commons*, 2004-2019), Ireland (*Dáil*, 1970-2022), and the United Kingdom (*House of Commons*, 1970-2022).[4] Combined, our dataset records the text of a total of 6,228,367 speeches.[5]

With a median number of words per speech between 30 (Ireland) and 736 (Australia), the texts in this dataset are quite long.[6] We have thus segmented each speech into paragraphs. This resulted in a total of 12,583,126 speech paragraphs.[7]

We will add more pre-training datasets in future iterations of this project. In the domain of party communication, we plan to add election manifestos (Lehmann et al., 2022) and

[4] We have extracted the speeches from the official XML-format transcripts provided by these countries' Hansard or parliamentary services.

[5] Australia: 101,020; Canada: 408,125; Ireland: 2,281,536; and United Kingdom: 3,437,686.

[6] These differences exists because the structure of the XML files differs across countries.

[7] Australia: 1,290,946 paragraphs (median of 65 words/paragraph); Canada: 1,809,600 paragraphs (median of 45 words/paragraph); Ireland: 4,324,958 paragraphs (median of 48 words/paragraph); and United Kingdom: 5,157,622 paragraphs (median of 61 words/paragraph).

press releases (Erfort et al., 2021). In the domain of legal documents (e.g., bills), we will extend our pre-training data by retrieving documents from ParaCrawl (Esplà et al., 2019), the Eur-Lex data (Ovádek, 2021), and publicly available records of the European Court of Human Rights.[8] And in the domain of international politics, we will integrate copora like the multi-UN corpus (Ziemski et al., 2016) and other datasets that record the proceedings of international organizations.

Further, we will extend the language coverage of our pre-training dataset. Multilingual quantitative text analysis is gaining in importance in comparative politics research (cf. Lucas et al., 2015; Licht, 2022), and many political text locations like the ParlSpeech2 dataset (Rauh & Schwalbach, 2020) or the Manifesto Corpus (Lehmann et al., 2022) record texts in many (European) languages.

3.1.2 Fine-tuning benchmarks

To evaluate our pre-trained models' performance in downstream tasks, we focus on supervised classification. Supervised classification is one of the most widely applied text analysis tasks in the political science literature (e.g., Hillard et al., 2008; D'Orazio et al., 2014; Burscher et al., 2015; Peterson & Spirling, 2018; Siegel et al., 2019; Barberá et al., 2021; Osnabrügge et al., 2021) and related fields (e.g., van Atteveldt et al., 2021; Bonikowski et al., 2022). Moreover, recent contributions indicate that supervised classification is one task where transfer learning might be especially beneficial (Linder et al., 2020; Laurer et al., 2022; Wankmüller, 2022; Licht, 2023). We have compiled a benchmark of currently two annotated text datasets available from the replication materials in highly-visible political science publications.

The first dataset included in our benchmark is based on the human-annotated text data collected by the *Comparative Manifestos Project* (CAP) (bau, 2019). The CAP coding scheme has 21 categories (so-called main topics), which we use as labels to fine-tune multi-class classifiers (cf. Linder et al., 2020; Laurer et al., 2022). We use the *Speeches from the Throne* dataset (1911-2012) compiled by (Jennings et al., 2011), which records 6624 labeled texts.[9].

The second dataset included in our benchmark is based on Benoit et al. (2016) who have distributed sentences sampled from British party manifestos for crowd-sourced human coding. Their coding task is a two-step conditional procedure. A coder first needs to decide whether the sentence they are asked to code discusses issues pertaining to a given poilicy area. If so, they were also aske to code the sentence's expressed stance on a 5-point scale.

Benoit et al. (2016) have adopted this coding procedure in two separate tasks. In their first task, they have focused on the policy areas of economic policy and social policy. The policy area codings for this task thus comprise three label classes ("economic", "social", and "other") and if the sentence was judged to belong to either the economic or social policy area, they also contributed stance codings (left to right for sentences with economic policy area focus, conservative to liberal for for sentences with social policy area focus). In their second task, they have focused on the policy areas of immigration. The policy area codings for this task thus comprise two label classes ("immigration" and "other") sentences that were judged to belong to the immigration policy area were further coded for their stance.

Their data thus provides us with labeled data to perform two classification tasks ("economic" vs. "social" vs. other and "immigration" vs. other) and three stance detection

[8] https://hudoc.echr.coe.int/

[9] see https://www.comparativeagendas.net/uk

tasks. Omitting the immigration policy stance detection task because there are very few sentences in this subset, we have prepared their data to construct a sentence-level labeled texts datasets.[10]

In future iterations of this project, we will add more labeled datasets to our benchmark to increase its diversity and coverage of relevant political text analysis tasks. For example, we will include existing pairwise comparison (Benoit et al., 2019; Carlson & Montgomery, 2017), multilabel (Theocharis et al., 2016; Erlich et al.), and stance detection (Bestvater & Monroe, 2022) datasets. We also plan to make our benchmark publicly available on GitHub.

3.2 Model architectures and training strategies

Our modeling approach is to use RoBERTa (Liu et al., 2019), a variation of BERT that has shown improved performance on a variety of natural language understanding tasks, to investigate the effect of additional pre-training on our domain-specific dataset and compare the performance of the resulting models against zero- and few-shot learning using OpenAI's GPT-3 API.

We first pre-train a RoBERTa model on a domain-relevant dataset consisting of English-language parliamentary speech records from the Irish Dáil. The pre-training is performed for 500 steps, where every step processes 100,000 documents. We create model checkpoints at steps 0, 50, 100, 150, 200, 250, and 500.

For each of these checkpoints, we load the pre-trained RoBERTa model and fine-tune it on the Benoit economic and social policy dataset, which is part of our benchmark dataset collection. The fine-tuning is performed for 10 epochs. By evaluating the models at various pre-training steps, we can assess the relative benefits of additional pre-training versus fine-tuning.

In parallel, we conduct an experiment using OpenAI's GPT-3 API to perform zero- and few-shot classification on the same benchmark datasets. This experiment allows us to compare the performance of our pre-trained and fine-tuned RoBERTa models with state-of-the-art few-shot learning techniques.

To employ few-shot learning, we provide GPT-3 with a series of examples, including a brief instruction of the task and several input-output pairs. These examples serve as context to help the model understand the specific classification task. To perform zero-shot learning, we provide GPT-3 with a single example, which is the instruction of the task. GPT-3 then generates a response for each new input, which represents the predicted class label.

3.3 Evaluation metrics and experimental design

To evaluate and compare the performance of the models trained using the three different strategies, we employ the following evaluation metrics.

Accuracy: This metric represents the proportion of correctly classified instances out of the total number of instances. It is a common and easily interpretable measure for classification tasks.

Precision, Recall, and F1-score: These metrics provide a more nuanced evaluation, considering both false positives and false negatives. Precision measures the proportion of true

[10] We have aggregated annotations of multiple crowd workers per sentence into single, sentence-level policy areas labels (stance codings) using majority voting (the mean) and used sentences' ground-truth labels (stance codings) whenever available.

positive instances among the instances predicted as positive, while recall measures the proportion of true positive instances among the actual positive instances. The F1-score is the harmonic mean of precision and recall, offering a single metric that balances both values. Because our validation task has three labels, we use F1 macro, which computes the unweighted mean of the F1-score for each class. That is, the F1 macro score is the average of the F1 scores for each class.

To ensure a fair comparison of the three training strategies, we divide our text-label pairs dataset into three subsets: training, validation, and test sets. The training set is used to fine-tune the BERT-like models, while the validation set is employed for hyperparameter tuning and model selection. The test set is reserved for evaluating the final models, providing an unbiased estimate of their performance.

4 Results

Figure 1 presents the preliminary results of our experiments, comparing the performance of RoBERTa classifiers with varying levels of domain-specific pre-training (denoted by the colored lines) and fine-tuning (denoted by the number of epochs) against zero- and few-shot queries using OpenAI's davinci-text-003 GPT-3 model (denoted by the grey and black dashed lines, respectively).[11]

There are four key takeaways from this figure. First, the RoBERTa classifiers consistently outperform GPT-3 zero-shot and few-shot queries across all levels of domain-specific pre-training and fine-tuning epochs. Second, the impact of domain-specific pre-training on the RoBERTa classifiers seems to be relatively minor: the performance difference between the colored lines (i.e., RoBERTa classifiers with different levels of domain-specific pre-training) is not substantial, particularly when the number of fine-tuning epochs is moderately high (e.g., 10 epochs).

Third, the performance of the RoBERTa classifier does not improve significantly with additional fine-tuning epochs, suggesting that the model reaches its optimal performance after only a few steps of fine-tuning. Fourth, (and most surprisingly), both GPT-3 approaches perform poorly, and the few-shot queries approach (black dashed line) shows only a marginal improvement over the zero-shot queries (grey dashed line), indicating that prompt engineering has limited effectiveness in enhancing the performance of GPT-3 in this specific domain.

We discuss these results in the next section.

5 Discussion

In this paper, we presented a preliminary investigation into the effectiveness of the domain-adaptation approach to pre-training and fine-tuning BERT models compared to zero-shot and in-context learning approaches using GPT models such as GPT-3. Our results showed that RoBERTa, when fine-tuned on the specific classification tasks, vastly outperforms both zero-shot and in-context learning approaches using GPT-3. While we need to conduct a more thorough series of tests, we suspect that this can be attributed to a combination of

[11] An epoch refers to one complete iteration of the training dataset during the fine-tuning process. A higher number of epochs indicates more exposure of the model to the task-specific labeled data, allowing it to learn more from the training set. Each pre-training step is a single update of the model's weights based on a batch of training samples. In our experiment, fifty training steps is equivalent to one hundred thousand speeches of legislative text observed by the model.

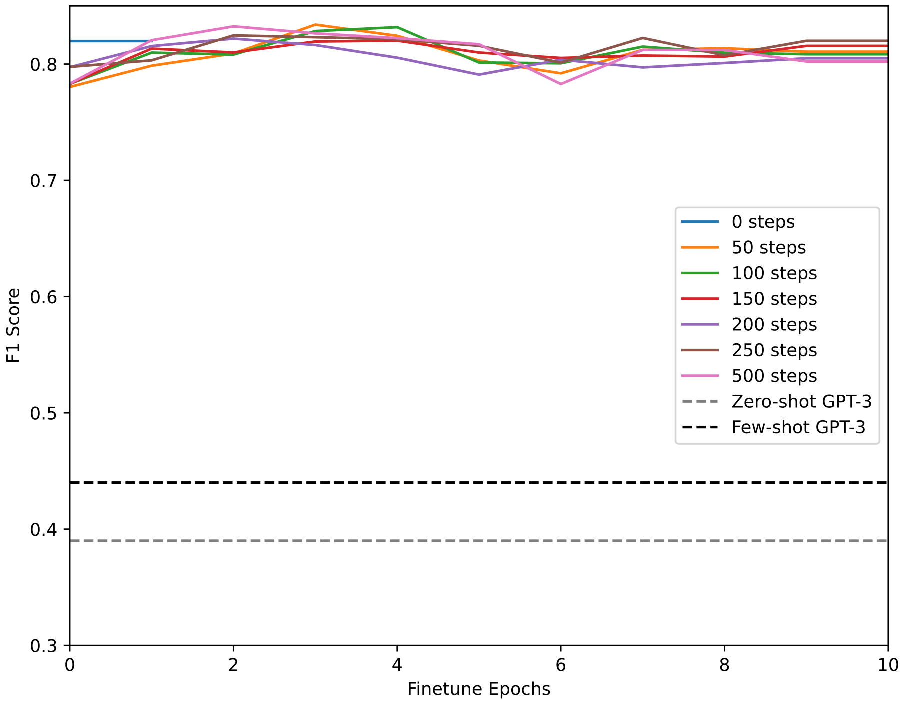

Figure 1: Comparison of RoBERTa classifier performance at varying levels of domain-specific pre-training (denoted by the colored lines) and fine-tuning (denoted by the number of epochs), with zero- and few-shot queries of OpenAI's davinci-text-003 GPT-3 model, denoted by the grey and black dashed lines, respectively.

RoBERTa's architecture, domain-specific pre-training, fine-tuning optimization, and task-specific supervision. The improved performance of RoBERTa demonstrates the benefits of fine-tuning on target tasks and utilizing task-specific labeled data.

We also showed evidence indicating that additional domain adaptation via pre-training does not seem to make a significant difference when there is a moderate amount of data (in our case, approximately 10,000 sentences in the training set) to train on. The initial pre-training of the RoBERTa model on a large corpus of general-domain text, then, might be sufficient to capture the necessary contextual information for the target tasks, making further domain-specific pre-training less impactful when a moderate amount of training data is available.

Finally, in our results prompt engineering only marginally improved the performance of the GPT-3 model. While more work is needed to determine exactly how in-context learning can improve the performance of GPT models in certain tasks, our results communicate a useful reminder that just because GPT models show a remarkable ability to generate text, they are not necessarily the best choice for all NLP tasks.

Our results suggest that while GPT-3 has demonstrated impressive capabilities in various tasks, it may not be ideally suited for certain types of classification problems, as generative nature and focus on language modeling may limit its ability to capture and discriminate fine-grained patterns in political texts, particularly when compared to a fine-tuned RoBERTa model that has been specifically optimized for such tasks.

We therefore recommend that researchers do not underestimate the continued power of fine-tuned BERT-family models such as RoBERTa for their classification tasks. Our findings suggest that this approach can yield significant performance improvements compared to zero-shot and in-context learning using GPT-3.

Further research is needed to confirm these findings and explore other factors that may contribute to the observed performance differences between the fine-tuning approach with RoBERTa and the zero-shot and in-context learning approaches using GPT-3.

To this end, we plan to expand our benchmark dataset collection, adding more labeled datasets that cover a broader range of political text analysis tasks, and use this data to investigate how different techniques of domain adaptation and in-context learning can be combined to further improve the performance of a wide variety of LLMs. We are particularly interested in exploring the domain adaptation of GPT models with political corpora using novel fine-tuning techniques, including multi-task learning and knowledge distillation, as well as in exploring creative applications of text generation models such as GPT-3 and GPT-4 for political text analysis tasks that extend beyond classification.

References

(2019). Comparative policy agendas: Theory, tools, data.

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *29*(1), 19–42.
URL `https://www.cambridge.org/core/product/identifier/S104719872000008X/type/journal_article`

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *110*(2), 278–295.
URL `https://www.cambridge.org/core/product/identifier/S0003055416000058/type/journal_article`

Benoit, K., Munger, K., & Spirling, A. (2019). Measuring and explaining political sophistication through textual complexity. *63*(2), 491–508.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12423`

Bestvater, S. E., & Monroe, B. L. (2022). Sentiment is not stance: Target-aware opinion classification for political text analysis. (pp. 1–22).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.

Bonikowski, B., Luo, Y., & Stuhler, O. (2022). Politics as usual? measuring populism, nationalism, and authoritarianism in u.s. presidential campaigns (1952–2020) with neural language models. *51*(4), 1721–1787.
URL `http://journals.sagepub.com/doi/10.1177/00491241221122317`

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020a). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.) *Advances in Neural Information Processing Systems*, vol. 33, (pp. 1877–1901). Curran Associates, Inc.
URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020b). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4.

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *659*(1), 122–131.
URL `http://journals.sagepub.com/doi/10.1177/0002716215569441`

Carlson, D., & Montgomery, J. M. (2017). A pairwise comparison framework for fast, flexible, and reliable human coding of political texts. *111*(4), 835–843.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.
URL `https://aclanthology.org/N19-1423`

D'Orazio, V., Landis, S. T., Palmer, G., & Schrodt, P. (2014). Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *22*(2), 224–242.
URL `https://www.cambridge.org/core/journals/political-analysis/article/`
`separating-the-wheat-from-the-chaff-applications-of-automated-document-classification-using-su`
`1E5431CC964E45218255584A4331D423`

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Erfort, C., Stötzer, L. F., & Klüver, H. (2021). Parties' evolving issue agendas.

Erlich, A., Dantas, S. G., Bagozzi, B. E., Berliner, D., & Palmer-Rubin, B. (????). Multi-label prediction for political text-as-data. *30*(4), 463–480.

Esplà, M., Forcada, M., Ramírez-Sánchez, G., & Hoang, H. (2019). ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, (pp. 118–119). Dublin, Ireland: European Association for Machine Translation.
URL `https://aclanthology.org/W19-6721`

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. In *IEEE Intelligent Systems*, vol. 24, (pp. 8–12). IEEE.

Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-assisted topic classification for mixed-methods social science research. *4*(4), 31–46.
URL `https://doi.org/10.1080/19331680801975367`

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Jennings, W., Bevan, S., & John, P. (2011). The agenda of british government: The speech from the throne, 1911-2008. *59*(1), 74–98. Publisher: SAGE Publications Ltd.
URL `https://doi.org/10.1111/j.1467-9248.2010.00859.x`

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, (pp. 1746–1751).

Laurer, M., Atteveldt, W. v., Casas, A., & Welbers, K. (2022). Less annotating, more classifying – addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. Publisher: OSF.
URL `https://osf.io/wqc86/`

Lehmann, P., Burst, T., Lewandowski, J., Matthieß, T., Regel, S., & Zehnter, L. (2022). Manifesto Corpus. Version 2022-2.

Licht, H. (2022). Cross-lingual classification of political texts using multilingual sentence embeddings. *0*(0), 1–14.

Licht, H. (2023). Cross-lingual classification of political texts using multilingual sentence embeddings. (pp. 1–14). Publisher: Cambridge University Press.
URL `https://www.cambridge.org/core/journals/political-analysis/article/crosslingual-classification-of-political-texts-using-multilingual-sentence-embeddings/30689C8798F097EEBA514ABE4891A71B`

Linder, F., Terechshenko, Z., Padmakumar, V., Liu, F., Nagler, J., Tucker, J. A., & Bonneau, R. (2020). A comparison of methods in political science text classification: Transfer learning language models for politics.

Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *23*(2), 254–277.
URL `https://doi.org/10.1093/pan/mpu019`

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, (pp. 3111–3119).

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, (pp. 11048–11064). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
URL `https://aclanthology.org/2022.emnlp-main.759`

OpenAI (2023). Gpt-4 technical report.

OpenSamizdat (2021). Chatgpt survey: Initial results. `http://opensamizdat.com/posts/chatgpt_survey/`.

Osnabrügge, M., Ash, E., & Morelli, M. (2021). Cross-domain supervised learning for topic classification of political texts. Tech. rep., Working Paper.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback.

Ovádek, M. (2021). Facilitating access to data on european union laws. *Political Research Exchange*, *3*(1), 1870150.
URL `https://doi.org/10.1080/2474736X.2020.1870150`

Pan, S., & Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, (pp. 1532–1543).

Peterson, A., & Spirling, A. (2018). Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems. *26*(1), 120–128. Publisher: Cambridge University Press.
URL `https://www.cambridge.org/core/journals/political-analysis/article/classification-accuracy-as-a-substantive-quantity-of-interest-measuring-polarization-in-westmi`
`45746D999CFCD1CB43E362392D7B2FB4`

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *arXiv preprint arXiv:1802.05365*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.
URL `https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf`

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer.

Rauh, C., & Schwalbach, J. (2020). The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies.
URL `https://doi.org/10.7910/DVN/L4OAKN`

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS 2019 Workshop on Machine Learning for Systems*.

Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., & Wang, L. (2023). Prompting gpt-3 to be reliable.

Siegel, A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., & Tucker, J. (2019). Trumping hate on twitter? online hate speech and white nationalist rhetoric in the 2016 us election campaign and its aftermath. Tech. rep., Working Paper.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil twitter use when interacting with party candidates. *66*(6), 1007–1031.
URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/jcom.12259`

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *15*(2), 121–140.
URL `https://www.tandfonline.com/doi/full/10.1080/19312458.2020.1869198`

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need.

Wankmüller, S. (2022). Introduction to neural transfer learning with transformers for social science text analysis. (p. 004912412211345).
URL `http://journals.sagepub.com/doi/10.1177/00491241221134527`

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning.

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, (pp. 3530–3534). European Language Resources Association (ELRA).
URL `https://aclanthology.org/L16-1561`