GeoPolitical Risk Predictor

Stanford CS224N Custom Project

William Denton
Department of Computer Science
Stanford University
wdenton@stanford.edu

Lucas Bosman
Department of Computer Science
Stanford University
lbosman@stanford.edu

Abstract

The geopolitical status of a country is at the core of all the functions of that country, such as its economy, military activity, and general populace sentiment. In this project, we are seeking to predict geopolitical risk by analyzing transcripts from congress, specifically for the United States. The hypothesis of this project is that geopolitical instability is reflected in the debates of congress, and as such we can use advances in large language models (LLMs) to analyze the significant dealings of these debates to predict instability. The challenges of this endeavor include reducing 100k daily tokens in a congress to a classifiable context size, appropriately fine-tuning an LLM with custom prompts, and extensive data-engineering to create the dataset. This applies work in parameter-efficient fine-tuning and summarising techniques to a novel challenge. We find promising results in using LLMs for this task and routes forward for future research in this topic.

1 Key Information to include

- External collaborators (if you have any): None
- External mentor (if you have any): None
- Sharing project: No
- Team Contributions: William created the dataset pipeline, as well as creating the fine-tuning pipeline, and writing the first draft of the approach, experiment, and analysis sections. Lucas found the data, cleaned it, created the congress transcript summarizing pipeline, and wrote the background and conclusion sections. Both worked on revising and editing the overall report.

2 Introduction

The effectiveness of a country's political system, its level of national security, and the growth of its economy are among the most important factors for determining both its present and future state. Not only this, but they are interrelated. When the government is in disarray, often then so is the economy, or that country's foreign affairs. Similarly, when a country is prosperous economically, typically that indicates a cohesive government due to less reason for division and disagreement. This aggregate measure of the internal and external opportunities and dangers of a given country, which we measure in geopolitical risk, is the measure that we wish to explore.
Just as geopolitical risk is multifaceted, so too are the ways one may measure it. Posts on Twitter or Facebook may provide a bottom-up view of the general sentiment of a country's populace, but ultimately these discussions are reflections or wishes for the actions of those with power or resources. Furthermore, the actions and words of congress have greater depth than such posts, as they reflect geopolitics via voting on laws, agreeing on political appointments, and debating contentious topics. Using congressional transcripts as a window into what's happening in government, then, is a perfect opportunity to gauge geopolitical risk. Yet, this information is often too overwhelming for individual

people to read through. As an anchoring point, the transcript from a single day's congress includes around 100,000 words. So, we sought to utilize advances in Large Language Models (LLMs) to parse through the extensive linguistic information present.
Current methods in predicting geopolitical risk include sourcing from newspaper headlines, and actively polling and crowd-sourcing from individual polls.Chatzis et al. (2018) Indeed, we use this former method as the predicted output for our own prediction task, but we hope that via advances in NLP our method could provide further insights, as the textual source is much more rich. Our approach of applying NLP to congressional transcripts for predicting congressional risk is novel. As a result, we face significant challenges: first, the dataset that we use is novel and has to be created from a variety of sources; second, we must somehow reduce the over 100,000 tokens present into a congressional transcript into a length which is usable for text classification; third, we must create a model which can accurately classify the reduced text into its category of geopolitical risk. Expounding on this final point, we fine-tuned LLaMA 2 on a classification task, building off established work that large pretrained models are well suited for fine-tuning on text classification. Li et al. (2023). We use parameter efficient fine-tuning techniques, specifically building off of the work of LoRA and using QLoRA in order to train our model on a limited compute budget. Dettmers et al. (2023)Hu et al. (2021).
The underlying hypothesis of this work is that there exists an observable correlation between congressional transcripts and geopolitical risk. This is the signal we seek to measure. In the case that we succeed, we hope to use our work to predict current geopolitical risk, which has abundant applications in foreign policy, finance, and beyond.Chatzis et al. (2018) We were able to produce promising results, especially given the difficulty of the underlying task, however in the future hope to build upon this work to increase the accuracy of our geopolitical risk measurement.

3 Related Work

Our work is at the intersection of two thriving areas of study, geopolitics and LLMs. Only in the last decade, and primarily the last five years, has geopolitics come to the fore of peoples' minds. The previous 40 years have been remarkably stable, as demonstrated by the Geopolitical Risk Index Caldara and Iacoviello. Additionally, LLMs have just emerged as a new area of study in Computer Science. Bar a few papers, such as the work by Chris Redl and Sandile Hlatshwayo on forecasting social unrest Redl and Hlatshwayo (2021) with traditional ML, few have applied AI to this area. We could not find anyone that had applied LLMs to predicting geopolitical risk. Instead, we discuss our influences from these areas separately.
In the field of Computer Science, employing LLMs for summarizing has been an increasingly important area of research. Currently, summary generation is being used by researchers Tam et al. (2023) to see if it is an appropriate benchmark for judging hallucinations. Across the field, multiple research teams are working on this sort of a metric. This implies two ideas. The first is that summarizing is an active area of research being built in part with LLMs. The second is that hallucinations are a problem in LLMs, particularly when generating summaries on inputted text. This informed our prompt engineering when we generated summaries. Specifically, we emphasized multiple times not to use outside information in its summary. Current techniques in automatic text summarization (ATS) include use of LLM's Jin et al. (2024) and techniques such as Retrieval Augmented Generation (RAG) Siriwardhana et al. (2023). Given limited context lengths of LLMs, longer documents pose a challenge for summarization. The common practical solution to this challenge is hierarchical summarization, where sections of a text are summarized, and a final summary is produced through feeding the previous summaries in.Jin et al. (2024) Due to the effectiveness of this solution, particularly for tasks like ours, we employ this method.
Furthermore, the main task for our project was that of natural language classification. To this end, there are several popular techniques, including regression, soft max end layers, and deep neural network classification layers built upon underlying LSTM, Transformer, or RNN embedding layers. Kowsari et al. (2019) However, many of these layers fail to leverage recent advances in pretrained modelling which have a better understanding of language than previously achieved. Thus, we sought to capitalize on the work of the Llama 2 team Touvron et al. (2023) and use a LLM for classification by adding a classification layer to the output of the text generated by the Llama 2 model, specifically using one which is fine-tuned for the classification task. This builds upon the work performed by Li et al. (2023). In performing this fine tuning, given our limited computing power, we made sure to use parameter efficient fine tuning (PEFT) techniques. Some popular techniques include adapters, Hu

et al. (2023), LoRA Hu et al. (2021), and pruning/subnetwork fine-tuning He et al. (2022). Building upon the work of He et al. (2022), we used a QLoRA based strategy which implemented many of the described techniques in the efficient adapatation created by Mangrulkar et al. (2022).
In politics and public discourse, geopolitics is of increasing importance. We see countries going to war in Europe and the Middle East, politicians making claims regarding a coming war between the US and China, and disruptions in global trade driven by attacks on global shipping lanes. In addition to this, there has been increasing instability domestically in the United States. Even in geographic science, researchers are giving consideration to what geopolitics means for the energy transition Yang et al. (2023). Given that these events and more indicate we are at a critical point for the future trajectory of the world, we felt that geopolitics was an important area to investigate.

4 Approach

Our approach for this project has been to distill the necessary congressional data from congressional transcripts, and then fine-tune Meta's Llama 2 (a LLM which is extensively pretrained) to predict geopolitical risk from the relevant congressional text. Touvron et al. (2023) In measuring geopolitical risk (GPR), we will use the measurement conducted by Caldara and Iacoviello which calculates a country's geopolitical risk on a given day based upon newspaper article headlines measuring geopolitical tensions. We thus have done extensive data-processing to first, transform this index into a classification problem and, second, match this GPR classification with the relevant congressional transcript text in order to create our dataset. Thus, this challenge has 2 main parts, creating the dataset on which we will fine-tune Llama 2, and then fine-tuning Llama 2 for classification.
In order to create our dataset, we are building upon a dataset created by Stanford investigators in which congressional meetings are transcribed, and additional data, such as word frequency, is computed. Gentzkow and Taddy (2018-01-16) The main challenge in constructing our dataset is in reducing these transcripts from over 100 thousand words in a day's congress, to at most 4096 (the maximum context window of Llama 2). In practice, we've found that the model has difficulty understanding the full context of 4096 tokens, especially given that (because of compute constraints) we use the 7B parameter Llama 2 model, which has reduced functionality when compared to the 70B parameter model. With these congressional transcripts, one approach we have taken is in randomly sampling excerpts from each day of congress. An advantage of this approach is that it carries the tone of the original congressperson, while a disadvantage is it's incredibly difficult to get an overall understanding of the entire day's proceedings from just a couple thousand words in random samples (each of 100 words).
Another approach we are taking is in summarization. In order to handle the smaller context window of the Llama 2 model, we sought to sample excerpts from the transcripts and summarize them. In Congress, speakers typically discuss a single topic at length when they have the floor, whether this topic be a bill, an appointment, or a matter of national security. The idea behind our approach is that we can capture the essence of this block of the given day's transcript in much fewer words, as much of a congressperson's speech is repetitive and we hope to gather their sentiment, the facts, and the issue at hand. We used another instance of Llama 2 to generate summaries of various parts of the text, and then stitched them together. This allowed us to feed in more information from the transcript into our fine-tuned LLM. Due to limitations on our compute and time, we were only able to sample a few parts of each transcript to generate sub-summaries. However, even with this approach, we were able to increase the amount of information we can feed our fine-tuned LLM about a given day's transcript by 300-400%.
We then matched each days' textual data with a classification of whether the geopolitical risk on that day is "high", "average", or "low." We computed that classification by comparing the GPR value for that day against an exponential average for the past GPR. This reflects the unidirectional aspect of real-life time, and also the exponentially decreasing importance window of geopolitical risk (the previous day's geopolitical risk can be significantly more important than a week ago.) If the day's GPR measure is within 20% of the average, then we classify that day as average, otherwise if higher or lower than normal, we classify it as "high" or "low", respectively. The code necessary in this process is entirely original, and has included significant exploration. Diagram 1 includes a visual summary of this dataset generation pipeline.

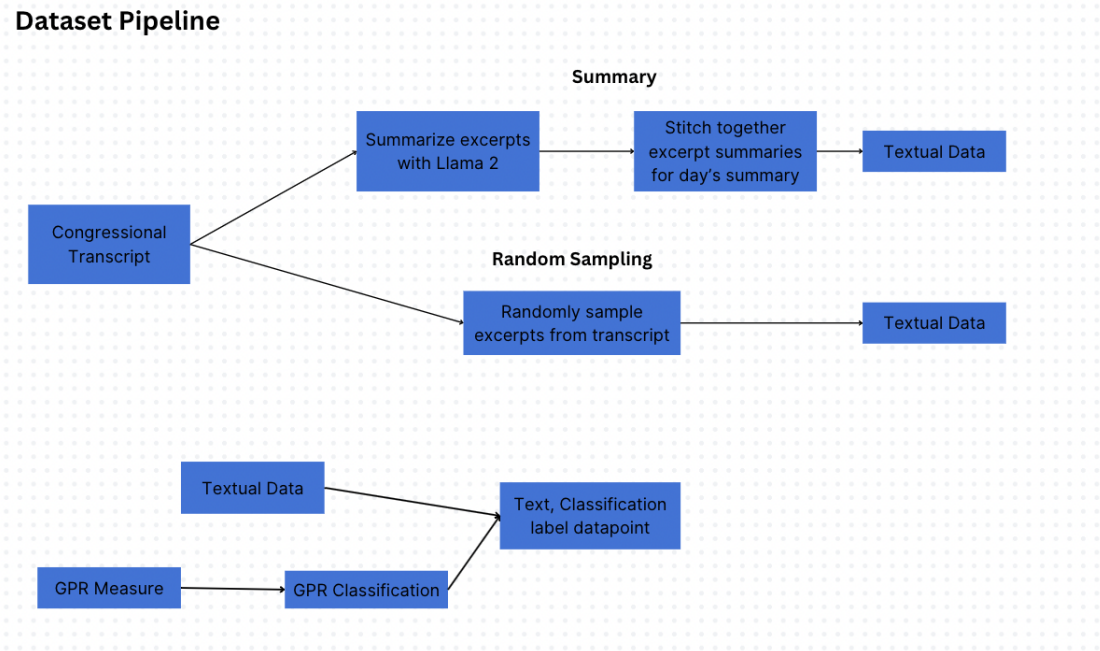

Diagram 1: Dataset Generation Pipeline

As a baseline, we are running Llama 2 without fine tuning on prompts engineered for classifying the congressional transcripts. For our model, we are fine-tuning Llama-2 on engineered prompts for classification, with the hypothesis that given the large textual abilities of Llama 2, it will come to learn patterns in congressional excerpts which are indicative of different levels of political risk. This is a state of the art practice, as the large pretrained models contain extensive abilities to understand language patterns, and by fine-tuning on specific congressional datapoints, we can leverage that general language understanding and apply it to our specific task of classification, teaching it how to recognize patterns in the transcripts which indicate different levels of political risk. In doing this fine tuning, we are combining reference code from a variety of resources, as well as writing code ourselves. Massaron (2024),Labonne (2023) Furthermore, since we are using an extremely large model, we are leveraging techniques in parameter efficient fine-tuning, specifically with QLoRA. Dettmers et al. (2023) Our fine-tuning parameters are mostly standard from reference sources as well as the original LoRA paper. Hu et al. (2021) Using this method, we can focus our limited computing power (especially given the incredibly large size of the model) on fewer parameters which will have the greatest impact to our classification task. Diagram 2 includes a visual summary of this model pipeline.

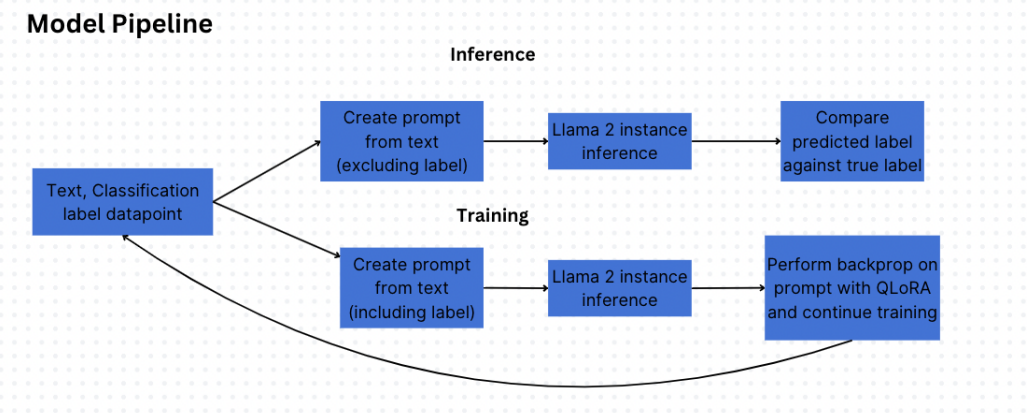

Diagram 2: Model Pipeline

5 Experiments

5.1 Data

As described in the previous section, our data is generated via a pipeline using the congressional transcripts generated by Gentzkow and Taddy (2018-01-16) and the GPR data created by Caldara and Iacoviello. Once we have classified each daily GPR datapoint into "high", "average", or "low" for a given day, we then match these classifications with a text prompt for that corresponding day.
For our text prompt, we tried 2 approaches. First, we attempted choosing random excerpts from the congressional transcript for that day since each day's transcripts is around 100,000 tokens, while the maximum context window length for Llama 2 is 4096 tokens. To this end, we tried two types of text data, one with 2000 tokens (more excerpts, but more difficult to get the attention mechanism to work correctly) and one with 600 tokens (fewer excerpts, but the language models are more easily able to understand that contextual size). We then performed additional prompt engineering to wrap these excerpts around instructions which tell the model to analyze these excerpts in order to classify the day's congressional hearing in terms of GPR.
The second type of text data we used was more sophisticated, and more succesful. For our second approach we summarized the transcripts using the process described in the approach section. We then matched the summary for a given day to its corresponding GPR index, and trained and tested the models on those datapoints.
For both of these types of text prompts, in order to actually fine-tune and evaluate the LLM, we had to do prompt engineering to wrap these text in prompts which instruct the model as to how to process them. We have slightly different prompts for the two types of textual inputs, but they are in most ways very similar (see Appendix A.1 for the prompts). They provide the context of the task to the model, and then ask the model to classify the corresponding text. In the case of evaluating the model, it's had extensive fine-tuning on classified text, and thus learns from the prompt to respond in one of the 3 possible classifications. This classification is the "output" for the model. In finetuning, however, the text that we train with includes both the prompt and the response, that is to say the output label is included in the fine-tuning text.

5.2 Evaluation method

In evaluating our model, we are using both a classification accuracy measure as well as a confusion matrix. This helps us to understand where the model is failing and where it's succeeding. Furthermore, for the most successful of the models that we trained and tested, namely the model which uses summarized transcripts instead of exerpts, we generated probability distribution plots. These plots have the probability range of each classification, and the frequency at which in the test set we observed that probability range, conditioned on the actual classification. The idea here is that for each underlying classification, we should see a shift to the right for the classification which matches the actual classification, as then our model actually learned useful trends in predicting the geopolitical risk.
In our accuracy and confusion matrix evaluations, we're seeking quantitative, objective measures of a models accuracy as well as it's ability to be heterogeneous with its predictions. With the distribution metric, we are seeking a quantitative measure which gives us a qualitative understanding of the ability of the model.

5.3 Experimental details

For our fine-tuning configurations, we used parameters described in the LoRA paper in addition to parameters provided in relevant guides. Massaron (2024),Labonne (2023), Hu et al. (2021) Specifically, we used a 4-bit quantization Dettmers et al. (2023), a LoRA dropout of 0.1, a LoRA alpha of 16, a learning rate of $2e-4$, a weight decay of 0.001, and fine-tuned for 3 epochs with about 1000 datapoints. We used the PEFT library Mangrulkar et al. (2022) for an efficient implementation of QLoRA. After completing training, we then merged our LoRA weights with the original 7B LLaMA 2 base model, and ran our inference.

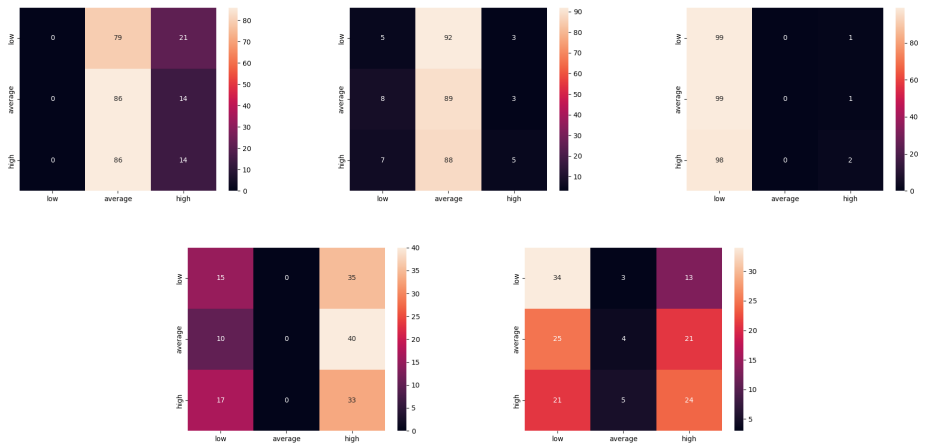

Figure 1: Confusion matrices for the baseline Llama 2 model evaluated on random excerpts (top left), a Llama 2 model fine-tuned and evaluated on 600 tokens of 200 token random excerpts (top middle), a Llama 2 model fine-tuned and evaluated on 2000 tokens of 200 token random excerpts (top right), the baseline Llama 2 model evaluated on summaries of congresses (bottom left), a fine-tuned Llama 2 model trained and evaluated on summaries of congresses (bottom right). In each figure, the row label represents the true label and the column label represents the predicted label.

5.4 Results

5.4.1 Accuracy

	baseline excerpts	600 context excerpts	2000 context excerpts	baseline summary	fine-tune summary
accuracy	0.33	0.330	0.337	0.320	.413

Table 1: Accuracy for several models, from left to right: the baseline Llama 2 model evaluated on random excerpts, a Llama 2 model fine-tuned and evaluated on 600 tokens of 200 token random excerpts, a Llama 2 model fine-tuned and evaluated on 2000 tokens of 200 token random excerpts, the baseline Llama 2 model evaluated on summaries of congresses, a fine-tuned Llama 2 model trained and evaluated on summaries of congresses.

The baseline models performed about as expected, equal to just random guessing. The fine-tuned models on a random collection of excerpts also performed similarly poorly to randomly guessing. This is because by randomly sampling 600 or 2000 tokens from over 100,000 tokens we aren't left with enough relevant context about the underlying events and major information. The model trained and evaluated on summarized meetings performed by far the best, with an almost 30% increase in accuracy relative to the baseline model.

5.4.2 Confusion Matrices

From Figure 1, clearly the baseline, 600 word context and 2000 word context models all have an extreme bias to predict a single classification for random samples. This is likely because there is a slight bias from the pretraining towards one of those classification words. After fine-tuning, the probability for all three classifications becomes extremely close, however there is a significant lack of signal in the random excerpts, so the underlying distribution for the next word, that is to say the classification, isn't effected, except possibly under extreme circumstances. Thus it almost always predicts the same word. Inspecting an example random excerpt prompt, this is also clear, as it is incredibly difficult to get any sense of the geopolitical risk for that day since it's such a small window into the day's dealings. Thus, although we sought to gain insight into sentiment of debates via the random excerpts, in order to predict geopolitical risk from when the debates are more heated or calm, we found that the model is much more successful when evaluating a summary of the debates, as the information of the congress is more important than the tones of the debates. This is reflected in the

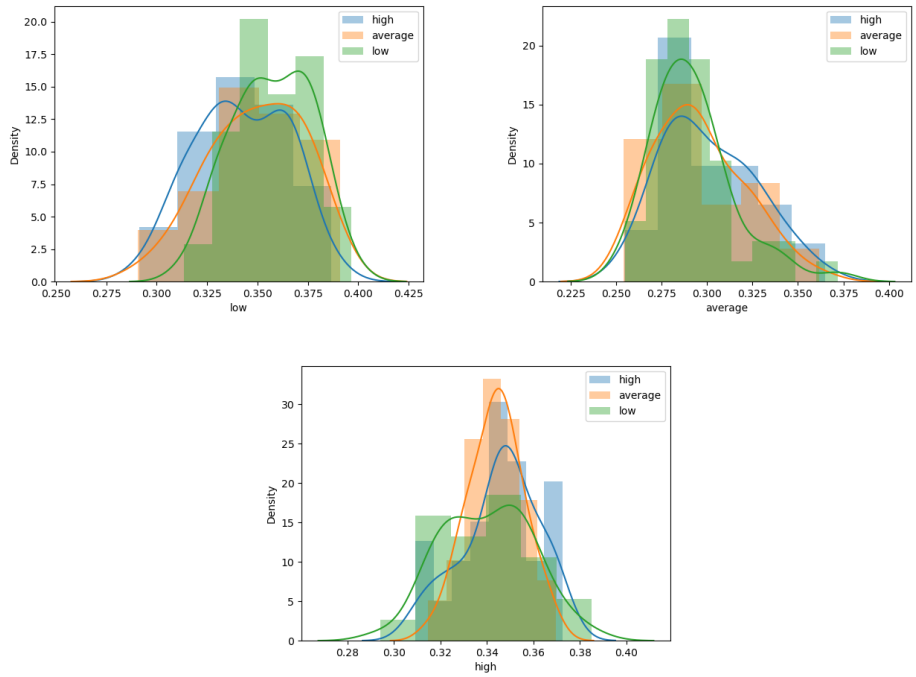

Figure 2: Classification probability (x) vs frequency (y) for each classification conditioned on the underlying label low (top left), average (top right), or high (bottom).

higher accuracy for the summary-based fine tune model as well as its more heterogeneous distribution in the confusion matrix. The failure of the excerpt model could also reflect the view of the author of this paper that politicians often disagree heatedly, regardless of the stakes.

5.4.3 Distribution of Fine-Tuned Summary Model

In Figure 2, we can see evidence that our summary model did learn insights into the geopolitical risk classification based upon the distribution shift. For the datapoints classified as low, we see a clear shift right for the green line and boxes, indicating that the low prediction is the most frequent, which is reflected in the confusion matrix. Although less clear, this same trend is evident in the datapoints conditioned on the average classification. The average probability is clearly shifted right when compared to the other graphs, at least relative to the low classification. Average is rarely the largest probability though, which is reflected in the lack of average datapoints classified as average. Lastly, when conditioned on being classified high, the high probabilities have a clear shift right, implying our model is learning that the high classification should be more likely when that is the correct classification.

6 Analysis

One of the main trends from our experiments is that the classification models which have random samples from the congress as their inputs fail. We hypothesize that the reason for this is quite simple, there just is not enough signal in that data. In trying to represent a 100,000 word context in about 2000 or 600 tokens, which are randomly sampled from that entire context, we lose too much information. The random excerpts may be irrelevant to the geopolitical risk, or even misleading. Qualitatively, when the authors tried to predict geopolitical risk from these excerpts, they found it extremely difficult because the congress is not adequately represented in those random quotes. It was our hope and intention that the models would be able to extract sentiment indicative of the geopolitial state of the congress from the tones of the congressmen and congresswomen, however this clearly pales in comparison to the lack of informational insight that comes as a result of such a relatively small set of

random samples.
The summary-based fine-tuned model performed better than we expected but worse than we hoped. We thought it could be entirely possible that there simply isn't enough of a signal in congressional meetings to indicate when there is geopolitical risk. The higher accuracy of the model, as well of our analysis of the classification distributions, indicates that the model did in fact learn valuable and insightful information for classifying congressional summaries. It achieved almost 70% accuracy on meetings of low risk, and almost 50% accuracy on meetings of high risk, which is significantly better than random guessing. In particular, however, it struggles on meetings of average geopolitical risk. This is because it's significantly easier to identify events which may cause risk to increase (war, financial issues, etc.) or decrease (good economy, peace, etc.), than it is to identify events which will cause geopolitical risk to remain as it is. In addition, congressional speeches or discussions may be inflamed or discussing issues in the language of high risk, even when there is a low or normal level of risk. For example, we may be at a low risk moment historically, but congressmembers speak as though we are at a level of higher risk. By inspecting datapoints with the average label which were predicted to be low or high risk by the model, we further justify this hypothesis. Often the congressional meetings are extreme in one way or another, rarely do they speak as if everything is just the same as it was, events are either very good or very bad. Furthermore, by translating a regressive measure into a classification output, we definitely lost some amount of information, leading to an imperfect classification labelling scheme. In context with this, we believe that our model actually performed quite well, and are excited by the possible future applications and development of this model.

7 Conclusion

Using LLMs to predict geopolitical risk was an enormous task. From creating the dataset to dealing with context window limitations to fine-tuning the large pretrained model, we overcame many obstacles. While our model did not perform at superbly high levels of accuracy, we feel we made considerable progress on this task, especially considering the difficulty of the underlying classification task. That is to say, as humans this task would still be extremely difficult, and thus there is no clear benchmark of how an extremely successful model would perform. We were able to start reading positive results from the model, significantly better than our baseline performance. With more compute and time, we are hopeful that we could continue improve our accuracy and prediction heterogeneity.
While we attempted to remove LLM hallucination through emphasizing excerpts from the transcript in our prompts, another route we could explore is the use of Retrieval Augmented Generation (RAG). RAG forces the LLM to retrieve relevant information from authoritative, pre-determined knowledge sources, which in this case would be the transcripts themselves. With this technique, we could combine the advantage of the summarizing technique of Llama 2: we were able to include more relevant information from the meeting; with the original reason for using quotations from the congress: including sentiment and moments of important interactions.
One avenue which may provide improved results is widening the scope of our data to include periods such as World War 2 or World War 1. These are prolonged periods of high geopolitical risk, and we think that this may give the model a stronger understanding of what it truly means to be at a high level. In turn, this may improve the classification of average risk. The high levels put into perspective how risks are being discussed in congress and what they are talking about at the extreme. From here, the model may be able to learn that moments of average risk are discussed differently, as in hypotheticals, as compared to actual events taking place. Our dataset of congressional transcripts starts with the 1980s, however, so a new or expanded dataset would be required to augment our analysis.
Furthermore, additional work could be done to explore how the classification scheme could be improved. We translated the continuous GPR values into a GPR classification for a given day, as described in our approach section, however there are many possible potential ways of creating the classification task from the continuous data, and some may show more consistent results.
Finally, we are curious whether this can be used for geopolitical forecasting. Future work may include taking an aggregate of the last 6-12 months of geopolitical risks, in combination with some sort of NLP for analyzing the issues causing geopolitical risk, and predict the future month's level of geopolitical risk.

References

Dario Caldara and Matteo Iacoviello. Geopolitical risk (gpr) index.

Sotirios P. Chatzis, Vassilis Siakoulis, Anastasios Petropoulos, Evangelos Stavroulakis, and Nikos Vlachogiannakis. 2018. Forecasting stock market crisis events using deep and statistical machine learning techniques. volume 112, pages 353–371.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jesse M. Shapiro Gentzkow, Matthew and Matt Taddy. 2018-01-16. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4).

Maxime Labonne. 2023. Fine-tune your own llama 2 model in a colab notebook.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label supervised llama finetuning.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods.

Luca Massaron. 2024. Fine-tune llama 2 for sentiment analysis.

Chris Redl and Sandile Hlatshwayo. 2021. Forecasting social unrest: A machine learning approach.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Yu Yang, Siyou Xia, and Xiaoying Qian. 2023. Geopolitics of the energy transition. volume 33.

A Appendix

A.1 Prompts

In this section we include the prompts that we used for wrapping our data. In the training data, the classification variable was filled with the correct classification, and in the test set, the classification variable was removed, and inference was left to the model to predict the next word (that is, the classification).

A.1.1 Random Excerpt Prompt

<s>[INST] «SYS» You are a political classification machine. You will be given excerpts from US congress meetings and you will analyze the overall nature of the excerpts as being indicative of "low", "average", or "high" political risk. «/SYS» Analyze the sentiment and content of the quoted excerpts from US congress meetings in the enclosed square brackets.
[**<RANDOM-EXCERPTS>**]
Now, determine if these excerpts are indicative of the US political and geopolitical risk being high, average, or low, and return the answer as the corresponding sentiment label "high" or "average" or "low". These excerpts are classified as: [/INST] **<CLASSIFICATION>**

A.1.2 Summary Prompt

<s>[INST] «SYS» You are a political classification machine. You will be given a summary of a US congress meeting and you will analyze the overall nature of meeting as being indicative of "low", "average", or "high" political risk. «/SYS» Analyze the sentiment and content of the summarized US congress meeting in the enclosed square brackets.
[**<SUMMARY>**]
Now, determine if this summary is indicative of the US political and geopolitical risk being high, average, or low, and return the answer as the corresponding sentiment label, only one of either "high" or "average" or "low". This summary is classified as: [/INST] **<CLASSIFICATION>**

A.2 Example Summary

Here we include an example summary created by our summarizing model:
Iran has shorter range Scud missiles and longer range North Korean missiles called Shahab III that can reach Israel.
Iran has launched a satellite into orbit using a very long-range missile called the Safir, which can also be used to deorbit a warhead.
Iran has thousands of uranium cascades operating to refine uranium and is fueling the Bushehr reactor, which will produce plutonium soon.
The greatest emerging threat to the US and Israel is Iran's missile and fissile material production, linked with the other speeches of Iran's head of state.
The US administration's missile defense intentions are uncertain, and it has canceled plans to upgrade US missile defenses, cut funding for the US-Israel Arrow 3 missile defense system, and offered to include Russians in NATO's missile defenses.
The preamble of the treaty negotiated by the State Department preserves Russia's ability to attack the US, and the administration has canceled plans to deploy the GBI system to Poland, leaving a real defense system for a hoped-for one.
The amendment should go forward without affecting the treaty if it does not limit the ability of the US or Israel to defend themselves, and it is necessary to fulfill the treaty's assertions that it has no relation to defense.
The 21st century should focus on fewer ways for nations to attack the US or allies and greater means for democracies, especially the US, to defeat an attack in case of war.
Senator DEMINT, Senator THUNE, Senator SHAHEEN, Senator RISCH, and Senator SESSIONS are recognized to speak for 10, 10, 10, 30, and 30 minutes, respectively.
The administration has banned offshore drilling in the Gulf of Mexico, which will cost thousands of jobs in the region, particularly in Texas and Louisiana.
The administration has been blocking drilling permits and only recently allowed drilling to resume.
Cuba, Russia, and China are expected to drill offshore in the Gulf, while the US is sending money to

Brazil and Mexico to allow them to drill off their coasts.
The administration's actions have raised questions about their commitment to American energy companies and their loyalty to the US.
Beverly Buchheit, a longtime educator and community leader in St. Louis, is being recognized for her dedication to service and her impact on the public schools and community.
Manuel Arianes, a Mexican man in the US illegally, was involved in a shooting with Border Patrol agents last Tuesday near the Mexican-American border.
The letter signed by several members of Congress, including the speaker, was sent to the President in October, expressing concern about the lack of National Guard support on the Mexican-American border.
The border is 1,980 miles long and the President sent 1,300-1,400 National Guardsmen to the border for a short period of time.
There is a war going on on the Mexican-American border with drug dealers, thieves, and terrorists entering the US from Mexico.
The US is training local law enforcement and military in Central America to stop drug dealers from moving into places like Costa Rica.
The administration's actions have been criticized for not addressing the problem of illegal immigration and drug trafficking on the Mexican-American border.