Politricks: Creating and Evaluating Artificial Political Speeches Using Large Language Models

Project Report

Jakob Amann Andri Rutschmann Elias Heppner

Abstract

As Large Language Models have AAAI creates proceedings, working notes, and technical reports directly from electronic source furnished by the authors. To ensure that all papers in the publication have a uniform appearance, authors must adhere to the following

instructions.

Introduction

The abuse of artificial intelligence for political gain is accelerating at an alarming rate. In light of scandals like Cambridge Analytica, vigilance against attacks on democracy is of utmost importance. Literature about the influence of fake news, social media, and bot networks on people's political behavior is abundant (Boxell, Gentzkow, and Shapiro 2017, 2018; Fujiwara, Müller, and Schwarz 2023). New technologies like microtargeting and deep fakes are already used as strategies to sway public opinion and manipulate voters at the ballots (QUOTE; QUOTE). The future potential of these and other technologies is hard to grasp and demands political regulation informed by rigorous research.

This project aims to contribute to this research by investigating the potential of generative Large Language Models (LLM) in emulating political content. The latest generative models have reached levels of yet unseen quality to the point of their output being virtually indistinguishable from real human-written content (see Biever, 2023). This content could easily be abused to change public perception about individual political figures or parties and would further accelerate existing manipulation strategies grounded in flooding social media with content.

Specifically, the present report seeks to find out how well OpenAI’s state of the art GPT-4 understands and emulates the political stance of different German parliamentary politicians concerning three contested political topics: Migration and Refugees, Covid-19, and the war in Ukraine. Thereby, we seek to contribute to the understanding of the model and its current abuse potential.

Beyond merely investigating generative capabilities, we also aimed to understand the ability of LLMs to capture political positions. Thus, we compare the generated content to actual speeches using a fine-tuned XLM-RoBERTa model, known as manifestoberta, from the Manifesto Project, which was trained on a large amount of annotated statements from political manifestos (QUOTE). As this model is limited to the dimensions specified in the Manifesto Project, we also implemented a principal component analysis (PCA) based approach running over the embedded speeches, which allows us to evaluate the LLM’s performance without any external constraints.

We prompted GPT-4 for speeches from a selection of German parliamentary politicians from all parties currently represented in the national parliament: AfD, CDU/CSU, FDP, Bündnis90/Die Grünen, SPD, and DIE LINKE, asking it to mimic the style and political stance of individual politicians and parties regarding the topics of interest.

This report proceeds by explaining the choice of data and preprocessing steps made. Following, the report explains the prompting and analysis strategy. The results of the two analysis approaches are presented and discussed with regards to the research question, limitations, the quality of the generated speeches, the applicability of the analysis methods, the abusive potential and avenues for future research. The project conceptualization was done by all team members, as we as the data preprocessing. The prompting was implemented by Andri Rutschmann, the manifestoberta analysis by Jakob Amann, and the principal component analysis by Elias Heppner.

Data Selection

We used a data set published by Open Discourse (Richter et al. 2020) containing the transcriptions of every plenary session of the German Parliament since 1949, totaling over 800,000 spoken contributions.

We chose the three topics of migration and refugees, Covid-19, and the war in Ukraine following an analysis of the most frequent keywords in a set of German news media articles from 2020 to 2022 (QUOTE). Among other topics, these three were ranked very high, and are commonly known as topics of discontent and public discourse. They are thus expected to have a high speech frequency alongside variance in politician and party stances necessary for our analysis.

Speech Selection

We chose to only investigate speeches given in the last three electoral terms (2013 - 2022). We assumed that this time frame yields a good trade-off between the size of the speech corpus and the variation in positions of parties and individual politicians. Choosing a too large time frame would make internal variation (party and politician) too large. Since we emulated speeches given by politicians, we also removed all speeches given by guests and secretaries of state. This resulted in a set of 121,001 observations.

After removing all non-politicians from the dataset we filtered by actual speech content. For the purpose of this analysis, we defined a speech as a standalone piece of spoken opinion. A question or interjection voiced as the result of some other contribution, for example, only makes sense in the context of the preceding contribution. We would not classify this as a speech. On the other hand, a contribution that addresses the whole parliament, reflects the opinions and stance of an individual politician, and stands on its own without any context, is considered a speech. For the selection process, we reasoned to rather be too restrictive and commit a type two error, as we want the analysis methods to capture semantic differences as much as possible, not other factors like topic, document type or length.

Most of the contributions start by greeting the parliament members and the president of parliament and are rather lengthy in terms of word count. Using this pattern, we constructed a regular expression (regex) filter that checks for different well-established forms of greeting within the first one to two sentences of each contribution. We supplemented our regex by investigating the most common tokens in the first few sentences of all contributions and then manually added some of them to the filter. A very common greeting, for example, is the use of “Kollegen” and “Kolleginnen” (colleagues). Using this regex filter we removed about 30% of all initial speeches, shrinking the data set to 83,848 observations.

Manually investigating the contributions revealed shorter contributions to be mostly responses or announcements. Therefore, we implemented a minimum token cutoff of 300 tokens and were satisfied by a check of a random sample of the remaining contributions This more than halved the entries in our data set from 83,848 to 36,053 speeches.

To further refine our selection of speeches, we implemented a regex-based topic selection, which identified speeches containing relevant words for the respective topic. After applying this filtering technique, we ended up with 3094 speeches about the Ukraine war, 2913 speeches about Covid-19 and 4247 speeches about migration and refugees. This filter was purposely defined to be rather broad, as our interaction with these texts made us realize that real speeches in the German Bundestag tend to be revolving around quite specific decisions, votes and laws, which certainly touch on our specified topics, but do not focus on them exclusively. Additionally, we assessed our methods to be robust against only partially relevant speeches.

Speech Prompting

Politician Selection

To prompt GPT-4 for LLM-generated speeches in the style of individual politicians we first had to come up with a selection of politicians. Ideally, we wanted each party to be represented by male and female politicians with both well and less known individuals. By themselves, these requirements would be pretty easily met but we also required the individual politicians to have held at least a few speeches on each topic in parliament. Otherwise, comparing the LLM-generated speeches to real ones would not have been possible. We settled for a minimum of four speeches per topic. This way, each politician viable for sampling had to have held at least four speeches concerning migration and refugees, the pandemic, and the Ukraine war respectively. We chose four speeches per topic since it was quite hard to find sufficient politicians per party that gave more speeches per topic. Nevertheless, we felt that four speeches per topic would allow us to get a sufficient estimate of the politicians’ real opinions and stances.

Filtering for all politicians passing the threshold for each topic, we manually selected four politicians per party. Where possible, we sampled two male and two female politicians and also tried to select one rather well-known as well as one less known politician per gender. Even so, it was impossible to sample a gender equal set of politicians from each party. It seems that, at least with regards to our three topics of interest, the speakers in parliament for the AfD, CDU/CSU, and FDP are mostly male.

The resulting sample of politicians consists of 24 politicians. Of those, 16 are male and 8 are female. As mentioned above for the AfD, CDU/CSU, and FDP no female politicians met our sampling criteria except for one female CDU/CSU politician. A full overview of the sampled politicians can be found in Table 1.

INSERT TABLE 1 OR APPENDIX

Prompting Strategy

We used three different approaches to generate artificial speeches. First, we prompted the model for speeches on each of the three topics on a party level. Here we did not ask the model to mimic the style, stance, and opinions of an individual politician, but rather those of an entire party. Next, we prompted the model for speeches in the style of individual politician while also including the party of each politician in the prompt. Finally, we repeated the preceding step but removed the politicians’ party affiliation from the prompt.

We used this approach to correct for any influence a politician’s party affiliation might have on the generated output. By prompting for party level speeches first, we established a baseline for each individual party. By prompting for speeches reflecting the style, stance, and opinions of individual politicians next, we captured the effect of including this information on the output in comparison to the party level output. Finally, by also prompting the model for speeches by individual politicians without giving their parties in the prompt, we aimed to reveal possible mediating effects of party affiliation on stance. Including a politician’s party in the prompt might trigger the model to compose a speech closer to the party’s general stance and opinions as opposed to the politician’s own. Generated speeches for a politician that might be considered a party outlier might be less extreme and closer to general party opinion if the prompt actively includes the politician’s party. The exact wordings of our prompts as well as the English translations can be found on GitHub. An exemplary German prompt looks like this:

*Schreibe eine Rede für den Bundestag im Stil von {politician} ({party}) zum Thema: "{topic}". Versuche dabei möglichst präzise den Standpunkt und die Ansichten von {politician} ({party}) zum Thema "{topic}" wiederzugeben und schreibe die Rede so, wie du erwarten würdest, dass sie vor versammeltem Bundestag persönlich gehalten werden würde. Bitte gib ausschließlich den Inhalt der Rede zurück. Falls du etwas ausgeben möchtest, was nicht Teil der Rede ist, fasse es in eckige Klammern "[]".*

There were no guardrail responses. In a few cases the model would append an explanation or disclaimer in squared brackets to the output. Due to this we were able to easily remove this non-speech content preceding the speech analysis.

The model returned four speeches per politician and topic giving us 288 speeches (24 politicians x 3 topics x 4 runs) respectively for the prompts containing the name and party of a politician as well as for the prompts containing only the name of the politician. Another 72 speeches (6 parties x 3 topics x 4 runs) were generated for the party level prompts not containing the name of an individual politician.

To assess the quality of the LLM-generated speeches we manually checked a random sample of speeches. The quality of these samples was astounding. Every generated speech was well written and did not contain any noticeable flaws. Just by going over these samples it was apparent that the model clearly varied the stances and opinions on the topics depending on which politician or party it was supposed to imitate. The model also captured the general style of German parliamentary speeches well, using similar greetings at the start of each speech and concluding with thanking the audience for its attention. All of these patterns can also be observed in the real speeches. This gives rise to the assumption that the model was trained with information on German parliament speeches.

Placement Analysis

We chose to apply two strongly differing methods to locate the speeches in dimensional spaces: A pre-trained, issue-specific BERT model, and an unspecific principal component analysis applied to the embedded speeches. Doing so, we can tell whether the results are driven by the placement method itself or rather driven by prompted speeches actually being accurate. Further, we are extending our contribution to not just investigating the generative, but also analytic capabilities of Large Language Models. Running an issue-specific and an unspecific model also enables us to test the robustness of both placement methods.

Manifestoberta

The Manifesto Project (QUOTE) focuses on the analysis of parties’ election manifestos in multiple languages in the period after World War II. As a part of this project, the manifestoberta model was trained and published. The data foundation for the model is a corpus of over 1.5 million human-annotated texts. Each text unit gets assigned one of the 56 categories, specified in the extensive Codebook. The XLM-RoBERTa model is designed to handle texts in a wide variety of languages, including German. The manifestoberta model is very well suited for our research interest as it was trained for exactly these kinds of classification tasks.

We decided to use the Sentence Model for our classification of the real and LLM-generated speeches. As this model has a maximum token length of 200 tokens, we decided to split each speech in individual sentences and aggregate the ratings for the 56 topics to get an average value for each topic for each individual speech. As our three chosen topics revolved around multiple different topics, we had to decide on a subset of dimensions for the final evaluation. Our final choice can be seen in Table 2. The definitions of these topics can be found in the coding handbook of the Manifesto Project. As the manifestoberta model returns numerical values for each dimension, it allows us to aggregate dimensions with positive and negative to one value.

INSERT TABLE 2

For the analysis, we further restricted our dataset on speeches from politicians selected for our simulation of their speeches using GPT-4. This resulted in a reduced data set of real speeches with 284 speeches about the Ukraine War, 225 speeches about Covid-19 and 464 speeches about Migration. Before we subsetted the speeches, we let manifestoberta annotate the whole data set and inspected the results visually. The subsetting yielded a solid representation of the overall data and was deemed to be a superior choice as our LLM-generated content was exclusively created using the personas found in these subsets, and therefore allowed a one-to-one comparison of real personas and their LLM-generated counterparts.

As we have an explorative research interest, our analysis of the results revolved around 3 pillars: visual exploration, comparing dimension averages and cosine similarity. The visual exploration, using party averages as well as individual values, focused on the detection of patterns in as well as between real and LLM-generated content. The comparison of the dimension averages for each person in each condition allowed us to estimate the extent to which the real and artificial speeches differ. As the visual exploration revealed that the artificial speeches tend to yield more extreme dimension values, we decided to use cosine similarity as a quantitative measure of similarity of vector orientation across all relevant dimensions for each topic.

Doc2Vec and Principal Component Analysis

The approach of mapping ideological placements by running a principal component analysis (PCA) on embedded political corpora was first established by Rheault and Cochrane (2020). Their contribution is a natural progression towards context embedding of previously often applied WordFish methods (QUOTE) and allows to incorporate control factors by tagging them onto the Doc2Vec encodings. They found principal component analysis of parliamentary corpora Doc2Vec encodings to be able to grasp the ideological variance in speeches in two - and multi-party systems, consistently clustering speeches by party.

We adopted their approach to our question of interest by encoding all speeches of a given topic, original and artificial, and running a PCA on the embeddings. Unlike the manifestoberta mapping, we kept all speeches on the topics as the embedding of each artificial and original speech hinges on the entirety of the corpus. The subset of speeches held by the prompted politicians would be too small. We chose to tag whether a speech was given by a member of government or not, as members of government have been shown to speak systematically different from opposition members (QUOTE). Thus, all speeches without a party assigned were omitted. The remaining dataset contained 2863 original speeches on the war in Ukraine, 3945 speeches on migration and refugees, and 2743 speeches on the Covid-19 pandemic. The Doc2Vec embedding was preceded by a common stop word removal, lemmatization, bigram and trigram phrase identification. We chose not to remove domain specific common words as they might carry semantic meaning (e.g. ‘Frau’ might be relevant in the context of migration, as the AfD frequently frames immigrants as a threat to women). We chose a rather large vector size of 300, a minimum word count of 50 words, 8 training epochs and a very large context window ±20.

We analyzed the results by plotting the totality of all speeches and the average positions per document type (original speech, artificial party speech, artificial politician speech, artificial politician-party speech). In addition, we investaged the average distance of artificial and original speeches to quantify the accuracy.

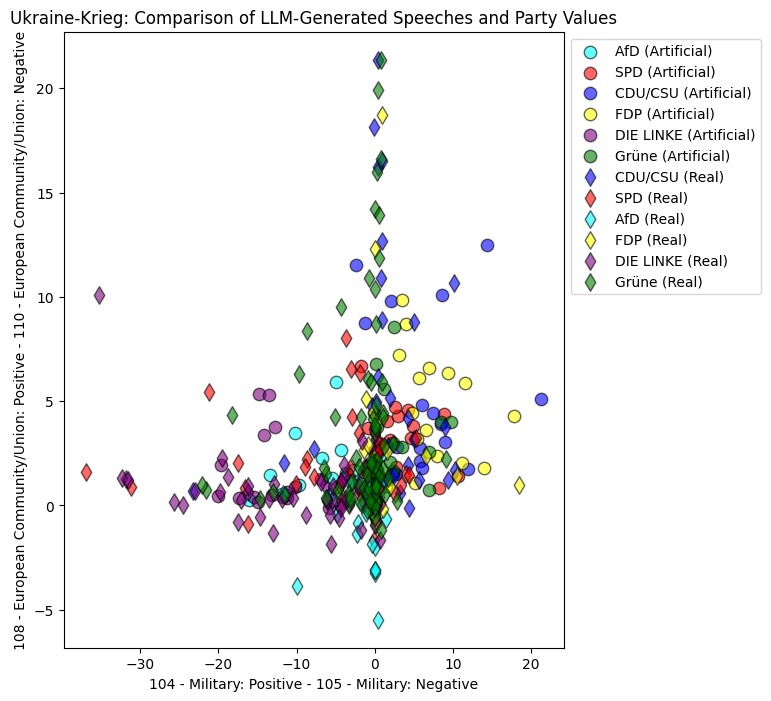
The results of the PCA must be taken with caution. The corpus is arguably very small, and similarity in doc2vec encodings is evidently strongly driven by word co-occurence and not semantic similarity (QUOTE). However, we are not just interested in the capacity of GPT-4 in generating speeches, but also in LLM’s capacity to accurately classify them. From our qualitative investigation, we know that GPT-4 was able to emulate political speeches rather accurately. Hence, we consider strong deviations because of the mapping process. To better understand the PCA, we also analyze the top 10 loaded words for each extreme and axis, allowing us to tell whether the PCA captured any political positions at all.

Results

Manifestoberta

The analysis of manifestoberta’s output concluded that the model is well-suited to map the political positions of texts accurately on the specified dimensions. It matched our evaluation of inspected samples as well as our overall expectations regarding typical political positions of certain parties or politicians. The most important findings during the analysis of the manifestoberta results revolved around two different aspects. First, the LLM-generated speeches tend to yield more extreme values in most of the dimensions than the real speeches, and, second, the LLM-generated speeches seem to be able to mimic the stance of the politicians relatively accurately.

To elaborate further on the first point, Figure 1 shows the rating of the individual speeches in the context of migration and refugees, colored by party.

We can clearly see that the AI-generated speeches (prompted with the name and party of the politician) have higher values on all shown dimensions. Overall, the relative as well as the absolute results are representative of the selected German parties. For example, we can see that the AfD is strongly positive regarding the idea of a National Way of Life and strongly negative regarding Multiculturalism. On the other hand, the left-wing party DIE LINKe is on the opposite side of the spectrum with some representatives of other left-leaning parties like SPD and Die Grünen. For the real speeches, we can see that they are way more centered, but the relative distribution (e.g., AfD and CDU/CSU positive National Way of Life and negative on Multiculturalism) seems to be as expected.

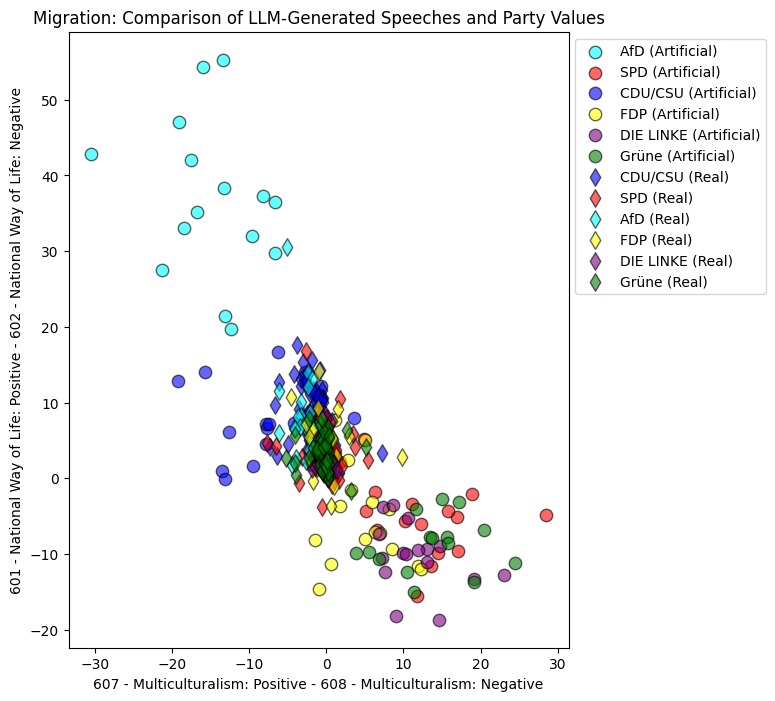
Figure 1: JAKOB ÜBERSCHRIFT

Figure 2: JAKOB ÜBERSCHRIFT

Regarding Figure 2 for the Ukraine War, we can see that the real speeches are actually more extreme in this case, except for the dimension Military Positive. Our expectations regarding the placement of the parties were met (e.g. DIE LINKE rated very negatively on the Military Dimension, AfD having the lowest values regarding the European Community/Union). This chart shows that not all artificial speeches are more extreme, but if we plot the averages, we can see that the LLM-generated speeches tend to be ranked higher or lower respectively.

Quantitatively comparing the average values of the real and the generated speeches per dimension for each politician per topic, we can confirm the impression we got from the plots. Most politicians have higher values for the generated speeches in most topics. This holds true for both conditions containing individual politicians in the prompt (with and without party).

As the higher absolute values for the LLM-generated speeches was clear from the visualizations and confirmed by comparing the averages for each politician, one could argue that GPT-4 failed in accurately mimicking the politicians. However, as mentioned earlier, the lesser focus of the real speeches on the topic of interest make it very plausible for the values to be lower. In order to evaluate the capability of GPT-4 to imitate the politician, we calculated the Cosine Similarity for the dimensions of interest defined by the respective topic between the manifestoberta values for the real speeches and the generated speeches for the conditions name only as well as name and party. The results revealed that the direction of the LLM-generated content was in line with the original speeches, often achieving values close to 1 (i.e., nearly perfect directional alignment). However, there are also some outliers with relatively low similarity values of around 0.3 for the name only condition. These politicians often achieve a higher similarity in the condition with the party included in the prompt. Overall, integrating the party in the prompt seems to be a very good strategy to avoid very poor similarities, but it also reduces the similarity in some cases compared to the name only condition. The cosine similarity between the artificial speech conditions (name with and without party) per politician was always very high, which is no surprise as the changes to the prompt were minimal and the politicians have likely appeared in the training data of the LLM with their respective party. Another interesting aspect of these findings is, that no politician had a low similarity across all topics and no topic had exclusively very high similarities, indicating that the LLM is able to capture the stance of politicians for multiple topics, but may not be able to capture the stance accurately for each topic, for each politician.

Doc2Vec and Principal Component Analysis

The principal component analysis of the embedded speeches yielded three important findings. Firstly, the analysis method does not locate artificial speeches close to actual speeches, neither for parties nor individual politicians. Secondly, it instead tends to cluster artificial and original speeches apart from each other, and artificial speeches rather close to each other. Thirdly, it does, to some extent, cluster speeches by party.

Regarding the imitation of political speeches, the results do not reveal a consistent pattern of close proximity between original speeches and prompted speeches. The average euclidean distances between original party and politician A diagram of a number of dots

Description automatically generatedspeeches and their artificial counterparts presented in Table 3 are lowest for the war in Ukraine, second lowest for migration and refugees and highest for Covid. across issues is XYZ, which is arguably large given the size of the PCA space. The averaged PCA values range between about minus one and one on the first dimension and zero and one on the second. Thus, these differences are rather large and show that the model does not approximate the artificial speeches close to the originals. This was confirmed further by qualitatively investigating some individual politicians.

A graph with different colored dots

Description automatically generatedTable 3

Figure 3: Average PCA Positions on the war in Ukraine

Figure 3 exemplifies the results for the topic of the war in Ukraine. The prompted speeches, i.e. politician speeches with or without party context, are not close to their original counterpart, yet closer to each other. They seem to be more extrem on both axes too. This is also true for the artificial party prompts not depicted above. The placing of the speeches does not seem to reflect an ideological space. Expectations concerning which parties should be far or close from each other are not met. When investigating the PCA loadings, it does show that semantically loaded, domain relevant keywords are present, yet they do not seem to be words strongly affiliated with a political leaning. An example can be seen in Table 4.

Table 4

A diagram of a number of dots

Description automatically generated

Figure 4: PCA Mapping of all Speeches on the Topic of Migration.

Concerning the overall clustering capacity of the approach, it shows that a PCA of domain-augmented, embedded political speeches is capable of clustering speeches from different parties to some extent. As shown in Figure 4, the clusters for the most part overlap, yet there evidently is a clustering of speeches.

Thus, overall, the PCA of document embeddings does not capture the close proximity of artificial speeches to original speeches found in the qualitative investigation of the outputs. The dimensions seem not to be driven primarily by ideology and semantic meaning. However, it does not produce arbitrary results either, and tends to cluster artificial speeches distinctly from actual speeches. Furthermore, though we cannot tell whether driven by semantic meaning, it at least is capable to detect party clusters to some extent.

Conclusion

With regards to our guiding research question, the findings of this research project indicate a strong performance of GPT-4 in mimicking stances of politicians across different parties and topics. Especially the qualitative analysis and manifestoberta yielded valuable insights into the models’ capabilities in reproducing political positions. Given the relevance, size and availability of the data set, we can conclude that the model was likely trained on the very speeches we used for the comparison. This could explain the relatively high cosine similarities.

Another significant finding is the great difference between the evaluation methods. While we were able to find a solid relation between the artificial and real speeches regarding their spatial mapping and their cosine similarity, we could not find a relevant relation between the real and artificial speeches using the embedded PCA approach. Thus, we conclude that task-specific models like manifestoberta are better suited for classification of political speeches. The PCA analysis was however able to identify clusters, both for parties and speech type (artificial vs. original).

Our findings highlight the potential of Large Language Models for classifying political content. Yet, the performance seems not to be good enough to replace manual classification, which is the current state of the art.

One of the main limitations, likely resulting in the strongly differing average values between the real and the LLM-generated speeches, especially using manifestoberta, was the difference in topic relevance. As we utilized a relatively broad topic filtering, most of the real speeches (sighted in various sampling steps) only touched on the specified topics of interest but did not deal with them exclusively over the course of the whole speech. This is likely due to the nature of the real speeches, as politicians discuss specific, rather granular laws, decisions or events in parliament and not broader, general topics like we asked the LLM to write speeches about. This difference in the groundtruth and generated data reduces the external validity of our findings. On the other hand, it opens up a very interesting opportunity for further research focusing on more granular topics (e.g., specific discussions in the Bundestag) and using generative LLMs to predict speeches of politicians for these topics.

The model chosen for generating our data, Open-AI’s GPT-4, is closed-source, and therefore could be deactivated or changed at any point in time, making the replication of our results potentially impossible in the future. Additionally, we can’t assess whether the LLM was trained on the speeches we used as ground truth.

Seeking to contribute to our understanding both about the generative and classification capabilities of LLMs, our project falls short of thoroughly establishing one or the other. Nevertheless, we deemed our overall contribution to the field as more relevant when investigating both aspects.

Our project opens up numerous pathways for future research and development. Developing classification models for political content could greatly increase the efficiency of classification and allow for the classification of much more data. Pertaining research, even within the scope of our project, adjusting model parameters, speech and politician selection would greatly add to the validity of the findings. Further, it would be beneficial to compare our findings with open-source and especially open-data models. Different prompting strategies could inform to what extent the generative capabilities depend on the quality of human input. Widening the scope of the analysis to different document types, such as interviews, different political settings, such as other systems, languages and countries, would solidify or nullify the robustness of the findings.

The results further raise concerns about the abusive potential of artificial intelligence. LLMs are indeed able to be abused to generate political content for malign use. Ongoing and future regulation processes should consider not only the potential of AI to affect individual rights, but it’s potential to undermine democratic institutions, manipulate public opinion and voting behaviours. Public discourse about how we want to live and interact with AI needs to incorporate considerations on political impacts at centerstage.

References

Biever, C. 2023. ChatGPT broke the Turing test – the race is on for new ways to assess AI. Nature 619, 686-689. doi.org/10.1038/d41586-023-02361-7.

Boxell, L.; Gentzkow, M.; Shapiro, J. M. 2017. Greater Internet Use Is Not Associated with Faster Growth in Political Polarization Among US Demographic Groups. Proceedings of the National Academy of Sciences 114(40), 10612-10617. doi.org/10.1073/pnas.1706588114.

Boxell, L.; Gentzkow, M.; Shapiro, J. M. 2018. A note on internet use and the 2016 U.S. presidential election outcome. PLoS ONE 13(7), e0199571. doi.org/10.1371/journal.pone.0199571.

Burst, T.; Lehmann, P.; Franzmann, S.; Al-Gaddooa, D.; Ivanusch, C.; Regel, S.; Riethmüller, F.; Weßels, B.; Zehnter, L. 2023. manifestoberta. Version 56topics.sentence.2023.1.1. Wissenschaftszentrum Berlin für Sozialforschung / Göttinger Institut für Demokratieforschung. doi.org/10.25522/manifesto.manifestoberta.56topics.sentence.2023.1.1.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. arXiv:1911.02116v2.

Fujiwara, T.; Müller, K.; Schwarz, C. 2023. The Effect of Social Media on Elections: Evidence from The United States. Journal of the European Economic Association, jvad058. doi.org/10.1093/jeea/jvad058.

Manifesto-Project. 2023. Manifesto Coding Instructions (5th re-revised edition). https://manifesto-project.wzb.eu/. Accessed: 2023-24-03.

Richter, F; Koch, P.; Franke, O.; Kraus, J.; Kuruc, F.; Thiem, A.; Högerl, J.; Heine, S.; Schöps, K. 2020. Open Discourse. Harvard Dataverse, V4. https://doi.org/10.7910/DVN/FIKIBO.

References

Dissertation or Thesis

*(Note: Include department and university):*

Clancey, W. J. 1979b. Transfer of Rule-Based Expertise through a Tutorial Dialogue. PhD dissertation, Department of Computer Science, Stanford University, Stanford, CA.

Forthcoming Book

Clancey, W. J. Forthcoming. *The Engineering of Qualitative* Models. Redwood City, CA: Addison-Wesley Publishing Company.

Preprint Server

Agrawal, A.; Batra, D.; and Parikh, D. 2016. Analyzing the Behavior of Visual Question Answering Models. arXiv preprint. arXiv:1606.07356v2 [cs.CL]. Ithaca, NY: Cornell University Library.

Published Book

Petroski, H. 1985. *To Engineer Is Human: The Role of Failure in Successful Design.* New York: St. Martin's Press.

Chapter in Published Book

Brown, J. S. 1977. Artificial Intelligence and Learning Strategies. In*Learning Strategies,* edited by J. O'Neil, 345–78. New York: Academic Press.

Forthcoming Journal Article

O'Connor, J. L. Forthcoming. Artificial Intelligence and Commonsense Reasoning. *AI Magazine*44(3).

Published Journal or Magazine Article

Cox, M. T. 2007. Perpetual Self-Aware Cognitive Agents. *AI Magazine*28(1): 32–45. doi.org/10.1609/aimag.v28i1.2027.

Paper Presented at Meeting

*(Note: Use this format only if no published proceedings appeared):*

Schoenfeld, A. H. 1981. Episodes and Executive Decisions in Mathematical Problem Solving. Paper presented at the 1981 AERA Annual Meeting. Boston, MA, September 24–30.

Zhou, S.; Suhr, A.; and Artzi, Y. 2017. Visual Reasoning with Natural Language. Paper presented at the AAAI 2017 Fall Symposium on Natural Communication for Human-Robot Collaboration. Arlington, VA, November 9–11.

Paper Presented at Meeting and Published in Proceedings

Lester, J.; Converse, S.; Kahler, S.; Barlow, T.; Stone, B.; and Bhogal, R. 1997. The Persona Effect: Affective Impact of Animated Pedagogical Agents. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems.New York: Association for Computing Machinery. doi.org/10.1145/258549. 258797.

Company Technical Report

Carbonell, J. R. 1970. Mixed-Initiative Man-Computer Instructional Dialogues, Technical Report QW-19871. Marina del Rey, CA: USC/Information Sciences Institute.

Scholarly Society Technical Report

Lin, F. 2007. Finitely-Verifiable Classes of Sentences. In *Logical Formalizations of Commonsense Reasoning: Papers from the 2007 AAAI Spring Symposium*. Technical Report SS-07-05. Palo Alto, CA: AAAI Press.

University Technical Report

Vattam, S.; Klenk, M.; Molineaux, M.; and Aha, D. W. 2013. Breadth of Approaches to Goal Reasoning: A Research Survey. In *Goal Reasoning: Papers from the ACS Workshop,*edited by D. W. Aha, M. T. Cox, and H. Muñoz-Avila. Technical Report CS-TR-5029. College Park, MD: University of Maryland, Department of Computer Science.

ArXiv Paper

Bouville, M. 2008. Crime and punishment in scientific re-

search. arXiv:0803.4058.

Website or online resource

NASA. 2015. Pluto: The ’Other’ Red Planet. https://www.

nasa.gov/nh/pluto-the-other-red-planet. Accessed: 2018-

12-06.