

Large Language Models for Measuring Contested and Multi-dimensional Concepts

Yaoyao Dai^{*,1} and Benjamin J. Radford^{2,3}

¹Department of Political Science and Public Administration, University of North Carolina Charlotte, Charlotte, NC, USA

²Public Policy Program, University of North Carolina Charlotte, Charlotte, NC, USA

³Joint first authors

*Corresponding author. Email: yaoyao.dai@charlotte.edu

Abstract

Large language models (LLMs) are deep neural network models pre-trained on massive amounts of text data. These models have demonstrated remarkable success in transferring linguistic knowledge learned in training to various downstream applications. However, one common concern when directly applying LLMs to the political science domain is the often complex and contested definitions of the target concepts, especially when the scholarly understanding of a concept may differ from the general audience writings that those LLMs were trained on. In this paper, we evaluate four zero-shot LLMs' performances in measuring populism, one of the most contested and popular concepts in academic and general public discussion in the past two decades. We compare this zero-shot approach with expert coding and a fine-tuned classifier. We find that ChatGPT 3.5 overall gives comparable results to expert coding and supervised learning. Predicting populism directly using LLMs can yield a measure with low face validity. However, with careful prompting, LLMs are able to provide measures better aligned with the scholarly understanding of a contested concept.

Keywords: Human annotation; measurement; biases; crowd-sourcing; text analysis

Introduction

While much of social science research relies on human-annotated data such as conflict events and regime types, high-quality human annotation can be time-consuming and expensive. Recent advancements in natural language processing, especially large language models (LLMs), provide a potential alternative to human annotations. LLMs are deep neural network models pre-trained on vast amounts of text data in a self-supervised fashion by, for example, masked language modeling or next sentence prediction. Once trained, those models can be directly applied to tasks that they are not trained on (zero-shot) or can be further fine-tuned for downstream applications. LLMs are found to have comparable or even better performance in various annotation tasks compared to human coders. For example, Törnberg (2023) find that ChatGPT4 achieved higher accuracy and reliability than Crowd workers in classifying the political affiliations of Twitter posters based on the content of tweets. Wu et al. (2023) demonstrate that ChatGPT gives stable and reliable answers in placing senators on the liberal-conservative spectrum in the pairwise comparison setting and those answers highly correlate with established measures of the senators' ideology. Palmer and Spirling (2023) find that LLMs can generate convincing and novel arguments about politics.

However, one common concern when directly applying LLMs to the social science domain is the often complex and contested definitions of relevant concepts, especially when the scholarly understanding of the concept differs from that of the general audience whose writings constitute the bulk of the LLM's training data. In other words, the vast contextual knowledge LLMs learn in the pre-training process might be a problem when this context diverges from the definition of the target concept. In this paper, we evaluate LLM performance in measuring such complex, subjective, and multi-dimensional concepts.

We test four LLMs' performance in measuring populism, one of the most contested and abstract concepts in social science. The rise of global populism in the past two decades also makes populism one of the hottest topics among the general public – the Cambridge Dictionary declared “populism” the word of the year 2017. Although a growing number of scholars across disciplines adopt the same ideational definition of populism, as a buzzword, populism is often poorly defined and used in the wrong way (Mudde and Kaltwasser 2018). In other words, while scholars have come to an agreement on the definition of populism, this understanding might be quite different from the vast amount of textual data LLMs are trained on.

Overall, we find that ChatGPT 3.5 gives comparable results to expert coding and supervised-learning, especially when we prompt the model to code sub-dimensions of populism. The three generative LLMs hosted on local computers (“local”) perform worse than ChatGPT, but still give reasonable-sounding answers to various prompts. The relatively smaller entailment model, BART MNLI, has a close performance to ChatGPT but at the cost of providing any “explanation” for its responses.

Data and Method

We replicate Dai and Kustov (2022) in measuring populist speeches in U.S. presidential campaigns and use their expert coding as the “ground truth.” We follow the ideational approach in conceptualizing populism as a set of ideas that portray society as being divided into two homogeneous and antagonistic groups, “the pure people” and “the corrupt elite,” and that emphasizes that politics should reflect the general will of the people (Mudde 2004; Mudde and Kaltwasser 2018; Hawkins and Kaltwasser 2018). Populism is a multi-dimensional concept that contains three necessary and jointly sufficient dimensions: people-centrism, anti-elitism, and anti-pluralism (Wuttke, Schimpf, and Schoen 2020). People-centrism specifies that people should be the only legitimate source of political power, a shared feature between populism and liberal democracy. However, populism is anti-pluralist and considers people as a homogeneous group with a general will and posits that politics is about finding and representing this (Müller 2017; Mudde and Kaltwasser 2017). Finally, populism is anti-elitist, and portrays elites as the opposite of the people and morally corrupt.

Dai and Kustov (2022) labeled 587 sub-speeches as either populist or not. Each such-speech is a sub-document from a presidential candidate’s campaign speech. 122 (20.78%) of those sub-speeches are labeled as populist. The original study applies a random forest classifier on locally-trained Doc2vec vectorization of the sub-speeches. We instead apply four zero-shot pre-trained LLM models to classify the sub-speeches, including one online model and three local models. In this section, we first explain the four LLMs and then discuss the procedures and prompts we use to classify sub-speeches according to the ideational definition of populism.

Online Language Models

We use OpenAI’s ChatGPT 3.5 (“ChatGPT”). We access ChatGPT via the commercial API service. Queries are submitted to OpenAI’s servers via web requests made in Python. The service charges by token, where a token represents a word or a piece of a word. At the time of writing, the cost was \$0.0015 per 1,000 prompt tokens and \$0.0020 per 1,000 response tokens. Our average cost across all prompts and responses was \$0.001 for a grand total of \$4.41 USD for 4,109 prompts and responses (excluding the ensemble experiment).

Because ChatGPT is operated as a service, OpenAI must have access to user submitted data (i.e., prompts) in order to process them. OpenAI’s privacy policy provides users the ability to allow their prompts to be used by OpenAI to improve their models, but this option is opt-in. The privacy policy also specifies that user prompts will be retained for up to 30 days, or longer depending on local laws and regulations. Researchers should be careful to verify that their research protocols allow them to share data with third parties prior to using the ChatGPT API. Fortunately, there are offline alternatives to ChatGPT that do not require any data to leave a researcher’s local computer.

Local Language Models

While ChatGPT is not available for download, several alternative language models are available for use on local computers. We use three different local language models. These have the dual advantages that they are free and that they do not require that sensitive data ever leave the researcher’s computer. The disadvantage is that these models are necessarily much smaller, in terms of total parameters, than ChatGPT. In our experiments, we chose only models that could fit on a single consumer-grade graphics processing unit (in our case, an Nvidia RTX 3090 with 24Gb of RAM). For comparison, ChatGPT 3.5 contains 175 billion parameters which, at four bytes each, would occupy several hundred gigabytes of memory alone.

These models are all pre-trained: they are LLMs that engineers have already fit to very large general-purpose text training corpora. Furthermore, they have all been additionally fine-tuned to task-specific datasets. We perform no additional fine-tuning of these models. The responses we elicit from the models are zero-shot in the sense that we have not provided any additional task-relevant information to the models other than that which is provided in the prompts themselves. We do not update the model parameters.

BART-large-MNLI: BART-large is a relatively small LLM with 400 million parameters (Lewis et al. 2020). BART MNLI has been pretrained on the Multi-Genre Natural Language Inference dataset (Williams, Nangia, and Bowman 2018). Natural language inference refers to the task in which a model is provided with a “premise” and a “hypothesis” and must return the probability that the premise entails the hypothesis. This can be used for zero-shot classification; by providing a text excerpt as the premise and a class label as the hypothesis, the model returns the probability that the hypothesized label describes the excerpt. Because this model is not a text generation model like the others evaluated, we modify our standardized prompts, given in Table 1, to accommodate the entailment format.

MPT-7b-instruct: The MosaicML Pretrained Transformer is a LLM that is permissively licensed (MosaicML NLP Team 2023).¹ It is licensed under the Create Commons Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0) license that is fully open source and requires only attribution and the transfer of the same license for derivatives of the model. This model comprises 7 billion parameters and has been pretrained on an instruction following dataset (Conover et al. 2023; Ganguli et al. 2022).

Falcon-7b-instruct: Falcon is a LLM published by the AI and Digital Science Research Center of the Technology Innovation Institute (TII) in Abu Dhabi.² This version, like MPT-7b-instruct, comprises 7 billion parameters and is pretrained on an instruction following dataset (Penedo et al. 2023). For both Falcon and MPT, we have adapted the prompts to follow the Dolly-15k dataset format as described in the MPT-7b-instruct model card (Conover et al. 2023).

Prompt Tuning

LLMs are known to be sensitive to their prompts; slight variations in the phrasing of a prompt can lead to dramatic differences in performance (Kojima et al. 2023; Liu et al. 2023). However, we perform no systematic prompt tuning in this project. Instead, we begin with the relevant definitions and condense them as much as possible such that they are legible, concise, and complete sentences. We then ask the model to perform its associated task in as few words as possible. We prefer brevity because some of the language models that we have chosen to evaluate have limits on the length of text inputs. Due to the length of the texts to be coded, we chose to minimize the length of the instructions.

1. <https://huggingface.co/mosaicml/mpt-7b-instruct>

2. <https://huggingface.co/tiiuae/falcon-7b-instruct>

Prompting

Our main prompts are given in Table 1. Since populism is a contested multi-dimensional concept, we design three kinds of prompts. In the first type of prompt, we do not define populism in the instruction. We expect the model to respond based on the learned meaning of populism in the pre-training stage. The simplest prompt is instructing the model to respond whether a supplied text contains populist rhetoric or not without giving the model the definition of populism (“populism short prompt”). Another simple prompt asks the model to answer the question as Cas Mudde, the highly cited populism scholar that defined populism in the ideational approach (“Mudde prompt”). We expect this prompt to result in closer codings to the expert’s codings, which also used Mudde’s definition, than the simple populism short prompt.

Table 1. Prompts used for predicting populist message

Populism long prompt	People-centric rhetoric claims that the people are the only legitimate source of political power. Anti-pluralist rhetoric claims that the people have a general will and that politicians should represent this will. Anti-elitist rhetoric claims that political elites are morally corrupt and intentionally subvert the people’s interests. Populist rhetoric is rhetoric that is people-centric, anti-pluralist, and anti-elitist. Does the supplied text contain populist rhetoric? First respond ‘YES’ or ‘NO’. Then, on a new line, explain your reasoning.
Populism short prompt	Does the supplied text contain populist rhetoric? First respond ‘YES’ or ‘NO’. Then, on a new line, explain your reasoning.
Mudde prompt	Pretend you are Cas Mudde. Does the supplied text contain populist rhetoric? Respond ‘YES’ or ‘NO’. Do not explain your reasoning.
Populism binary prompt	People-centric rhetoric claims that the people are the only legitimate source of political power. Anti-pluralist rhetoric claims that the people have a general will and that politicians should represent this will. Anti-elitist rhetoric claims that political elites are morally corrupt and intentionally subvert the people’s interests. Populist rhetoric is rhetoric that is people-centric, anti-pluralist, and anti-elitist. Does the supplied text contain populist rhetoric? Respond ‘YES’ or ‘NO’. Do not explain your reasoning.
Multi-dimensional prompts	
People-centrism	People-centric rhetoric claims that the people are the only legitimate source of political power. Is the following text people-centric? First respond ‘YES’ or ‘NO’. Then, on a new line, explain your reasoning.
Anti-pluralism	Anti-pluralist rhetoric claims that the people have a general will and that politicians should represent this will. Is the following text anti-pluralist? First respond ‘YES’ or ‘NO’. Then, on a new line, explain your reasoning.
Anti-elitism	Anti-elitist rhetoric claims that political elites are morally corrupt and intentionally subvert the people’s interests. Is the following text anti-elitist? First respond ‘YES’ or ‘NO’. Then, on a new line, explain your reasoning.

In the second type of prompt, we include Mudde’s abstract definition of populism in the prompt. In this “populism long prompt,” we give the model the definition of populism first, ask the model to respond “Yes” or “No,” and then to explain its reasoning. In the populism binary prompt, we give

the model the same definition of populism but do not ask the model to give its reasoning.

In the last type of prompt, we give the model the definition of each of the three core dimensions of populism and ask the model to code each dimension separately (“multi-dimensional prompts”). In addition to answering whether the supplied text contains the sub-dimensional feature, we ask the model to explain the reason behind the coding. While populism is a hot topic, each of the core dimensions of populism is much less common in popular writing. We thus expect this prompt to produce the closest coding to the experts using the same definition.

We attempted to minimize changes to these prompts when adapting them for the various models. When reframing our prompts for natural language inference, for use with `bart-large-mnli`, we use Huggingface’s default hypothesis template “This example is { },” with the curly braces replaced by our hypothesized rhetorical devices. We use the phrases “populism,” “anti-elitism,” “anti-pluralism,” and “people-centrism” without giving the model definitions of those phrases in some prompts, and use the definitions given in Table 1 in other prompts.

In reframing our prompts for the two instruct models, we follow the Dolly 15k format recommended by the `mpt-7b-instruct` model card:

Below is an instruction that describes a task.

Write a response that appropriately completes the request.

Instruction:

{Instructions}

Response:

We replace {Instructions} with the prompts shown in Table 1.

Evaluation

Table 2 provides the summary statistics of the LLMs’ performance on classifying populist speeches using Dai and Kustov (2022) coding as the “true” labels. We also include the model performance of the random forest model in the test set from Dai and Kustov (2022) for comparison.³ As we expected, the aggregated multi-dimensional prompting approach consistently performs the best

3. Dai and Kustov (2022) use active-learning to query and populate the training data. The test set is only a random sample of the initial seed training data. Given the nature of active-learning, the queried data are likely to comprise borderline or harder cases than the seed training and test data. Therefore, we include all models’ performance on the test set only in the Appendix.

within each model in terms of accuracy, where the model codes each dimension separately based on the prompt. A sub-speech is then coded as populist only if it contains all three core components of populism. Overall, ChatGPT 3.5 performs best in all prompts. The ensemble of five ChatGPT 3.5 responses to the populism long prompt performs equally well as a single ChatGPT model using the multi-dimensional prompt. The simple populism short prompt, which does not contain the scholarly definition of populism generally performs the worst with respect to accuracy and precision. It is also more likely to classify a sub-speech as populist.

Table 2. Model performance metrics for measuring populist speeches.

		Accuracy	Precision	Recall	F-1 Score
ChatGPT 3.5	Populism Long	0.66	0.35	0.73	0.47
	Populism Short	0.57	0.30	0.84	0.45
	Populism (Yes/No)	0.71	0.39	0.74	0.51
	Mudde	0.69	0.34	0.56	0.42
	Aggregated Multi-dimension	0.78	0.45	0.23	0.30
ChatGPT 3.5 Ensemble					
	Populism Long	0.74	0.43	0.70	0.53
MPT	Populism Long	0.48	0.23	0.66	0.34
	Populism Short	0.42	0.22	0.71	0.34
	Populism (Yes/No)	0.51	0.22	0.52	0.31
	Mudde	0.48	0.23	0.62	0.33
	Aggregated Multi-dimension	0.66	0.24	0.30	0.27
Falcon	Populism Long	0.51	0.21	0.51	0.30
	Populism Short	0.55	0.20	0.38	0.26
	Populism (Yes/No)	0.58	0.20	0.35	0.26
	Mudde	0.57	0.17	0.27	0.21
	Aggregated Multi-dimension	0.68	0.26	0.30	0.28
BART	Populism	0.47	0.27	0.89	0.41
	Three Dimensions	0.45	0.25	0.84	0.39
	Aggregated Multi-dimension	0.55	0.25	0.57	0.34
	Aggregated Multi-dimension Long	0.70	0.33	0.41	0.36
	Dai and Kustov (2022) supervised (test)	0.91	0.45	0.83	0.59

Figure 1 provides the correlation matrix of predictions from each prompt and model. Predictions from the ChatGPT ensemble model correlate most with the expert coding. On the other hand, predictions from MPT and Falcon have low correlations with the expert coding. Predictions from different prompts also have low correlations with each other in the same MPT and Falcon model.

We include two sub-speeches and responses from the online and local LLMs below. The first example is a random sub-speech that is coded as populist by the expert coders and the ChatGPT ensemble model. For the space limit, we omit parts of the input sub-speech in the examples. When

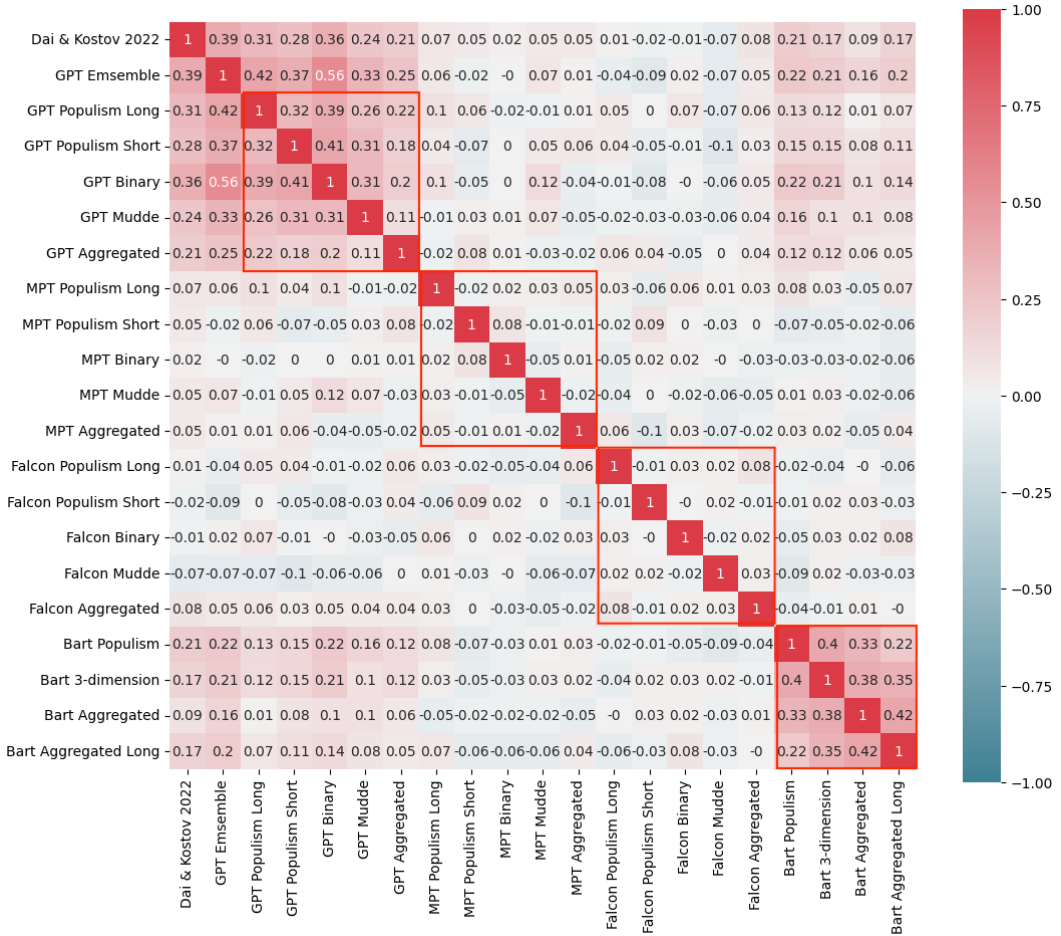


Figure 1. Correlation Matrices of Predictions

we instruct the LLMs to use the ideational definition of populism (populism long prompt), ChatGPT correctly identified the parts of the speech that reflect each of the core features of populism. While the local MPT model also correctly classifies the sub-speech as populist, the reasoning is not as relevant as is ChatGPT's. When we do not give the LLMs the scholarly definition (populism short prompt), ChatGPT's reasoning focuses on speech features that appeal to the ordinary people and criticize the elites, which lacks the necessary anti-pluralist feature of populism. Similarly, MPT's reasoning also focuses on appealing to the ordinary American, which is one common understanding of populism in the general public but lacks the core anti-pluralist feature of populism. It also includes features such as inclusion and equality that are irrelevant to populism in its reasoning.

True positive example

Input: “I promise you tonight as your president I will govern based on principle not poll numbers. I pledge to you that every day I will fight for the public interest not the corporate interests. I will not abandon any segment of American society — whether you’re gay or black or Latino or poor or working class...People are excited to be part of a political revolution that will change this nation and give us a future to believe in... Well my friends, let us make sure that everyone knows that on the issues of equality, and justice, and ending a rigged economy that is held in place by a corrupt political system, on battling climate change, on halting the draining of American jobs to faraway places, on these issues and so many more: We are in earnest; We will not equivocate; We will not excuse; We will not retreat a single inch ...” (Bernie Sanders, 2015)

ChatGPT response to populism long prompt: “Yes. The supplied text contains populist rhetoric because it is people-centric, anti-pluralist, and anti-elitist in nature. The statements emphasize the speaker’s commitment to representing the public interest over corporate interests, promising not to abandon any segment of society for political expediency. The text also highlights the need to build a coalition of voters that will bring about a political revolution and change the nation. The references to fighting against a corrupt political system and the mention of a ‘rigged economy’ reflect an anti-elitist sentiment. Finally, the repeated phrases of being in earnest, not equivocating, not excusing, and not retreating indicate a strong dedication to the people’s cause and a desire to be heard.”

ChatGPT response to populism short prompt: “Yes. The text contains populist rhetoric because it appeals to the interests and values of the ordinary people and criticizes the elite. The speaker promises to fight for the public interest, not the corporate interests, and pledges to not abandon any segment of American society. The text also emphasizes the need for a political revolution and building a winning coalition of voters. The use of strong and assertive language, such as ‘I will not retreat a single inch’ and ‘we will be heard,’ further reinforces the populist rhetoric.”

MPT response to populism long prompt: “Yes, the text contains populist rhetoric. There are multiple arguments that rely on populist assertions. Such as, The abolitionist William Lloyd Garrison wrote them when attacking the evils of slavery: I am in earnest—I will not equivocate—I will not excuse—I will not retreat a single inch—and I will be heard. This can be read as the speaker attacking those in power and asserting that the people will stand up and be heard. Here the speaker is claiming a moral superiority as their intentions come from a place of caring for the people rather than for personal gain or social status. I think this rhetorical strategy should be considered populist because it is putting the people’s interests above the elites’, and the people are the true source of political power.”

MPT response to populism short prompt: “Yes. The supplied text contains populist rhetoric as the opening statements clearly promote a presidential campaign based on a message of inclusion, equality, respect for all, and fighting for the interests of ordinary Americans. These are very common themes associated with populist rhetoric.”

The second example is a random sub-speech that the experts coded as not populist but predicted to be populist by the ChatGPT ensemble model. While ChatGPT gives the correct prediction

in the populism short prompt, the reasoning is not correct. ChatGPT argues that the text is not populist because it contains no explicit appeal to the emotions or desires of the people – a common but not defining feature of populism. Interestingly, although ChatGPT wrongly codes the message as anti-elitist in the populism long prompt, it gives the correct prediction and reasoning in the separate anti-elitism prompt. Although we give Falcon the scholarly definition of populism, its reasoning uses irrelevant “context knowledge” that is likely from training data. “Call for lower taxes” and “stronger borders” are the salient policy priorities of two prominent U.S. presidential candidates that are perceived as populists by the general public. When we do not include the definition of populism in the prompt, Falcon only gives a binary answer without reasoning.

False positive example

Input: “SACK: You’ve said in the past the plan could gravitate either way...EDWARDS: There is nothing backdoor about it...We’re going to let America decide what health care system works for them. I’m not deciding that for them. America’s going to decide it... The threshold question of any health care plan is ‘Is it universal?’ and his plan is not universal. So I don’t think it’s a minor thing. I think it’s a serious thing. He’s to be applauded for coming out with a serious plan, let’s be fair. But his plan is not universal ...I think there is a fundamental divide between my plan and Senator Clinton’s plan and his, because ours are universal and his is not. And I think we need universal care.”

ChatGPT response to populism long prompt: “YES. The supplied text contains populist rhetoric. The speaker, Edwards, argues that the people should have the power to decide on the healthcare system that works best for them. He emphasizes that he is not deciding for them and that they should be given the option to choose. This people-centric approach is characteristic of populist rhetoric. Additionally, Edwards dismisses the concerns of Republicans and positions himself as the candidate for serious and comprehensive healthcare reform, appealing to the idea that the political elites are not serving the interests of the people.”

ChatGPT response to populism short prompt: “No. The supplied text does not contain populist rhetoric. The conversation is focused on discussing different healthcare plans and their implications, as well as addressing potential criticisms and differences between the candidates’ proposals. There is no explicit appeal to the emotions or desires of the people, which is a common characteristic of populist rhetoric.”

ChatGPT response to anti-elitism dimension prompt: “No. This text is not necessarily anti-elitist. While it does criticize the healthcare plans of other politicians, it does not explicitly claim that political elites are morally corrupt or intentionally subverting the people’s interests. The focus of the text is primarily on the differences between the healthcare plans and the importance of achieving universal coverage.”

Falcon response to populism long prompt: “Yes. The text contains populist rhetoric when it uses the phrase ‘the people are the source of political power’ and when the

senator claims his plan will be ‘universal’ and ‘unquestionably more comprehensive’ due to his plan’s inclusion of a mandate. However, it does not contain other examples of populist-style rhetoric such as calls for lower taxes or stronger borders.”

Falcon response to populism short prompt: “yes”

Given the open-ended nature of LLM responses, these powerful tools can provide plausible and nuanced responses to queries. These may be helpful for researchers who require more than a simple binary value or class label. However, this flexibility also complicates the use of LLMs for researchers who do require a definite label: there are many ways in which an LLM can respond to even a simple question and researchers must account for these. This adds an additional step to the coding process when using LLMs; the LLM output must be parsed to identify the label. Sometimes this might be straightforward, but we found it often was not. We explicitly request ‘yes’ and ‘no’ responses from our models but found that some models inconsistently followed this instruction. ChatGPT consistently complied with our request for simple “yes” and “no” responses.

Conclusion

Large language models have demonstrated remarkable success in transferring linguistic knowledge learned in training to various downstream applications. However, the rich contextual knowledge can be both a blessing and a curse. When the target concept has different meanings in public writing and scholarly writing, the contextual knowledge might lead to a measure with low face validity. In this paper, we evaluate four LLMs in measuring one of the most contested and hardest-to-measure concepts: populism. We demonstrate that LLMs do draw irrelevant or wrong contextual knowledge of populism in reasoning when we do not provide the scholarly understanding of populism. However, with careful prompting, LLMs, especially ChatGPT, are able to align the reasoning more with the scholarly understanding.

We want to caution against taking the wrong message from this paper: while the supervised model with locally trained Doc2vec outperforms LLMs, we do not believe that document embeddings plus random forests are necessarily superior to the LLMs evaluated here. Instead, our goal was to explore the gap between zero-shot methods of coding long-form texts with LLMs and supervised methods for classifying those texts. Supervised learning, while typically cheaper than hand-coding, is still more costly than zero-shot classification. As LLMs proliferate, we expect that their use in zero-shot

settings for data coding (of all types) will grow rapidly. However, we find that current LLMs still lag behind custom supervised text coding solutions, with the possible exception of ChatGPT. Researchers should take care to consider their prompting strategies, select LLMs that are sufficiently powerful for their task, select a task formulation that leverages the characteristics of their task and their research constraints (e.g., natural language inference versus text generation), and to evaluate their models' outputs against trusted examples.

References

- Conover, Mike, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: introducing the world's first truly open instruction-tuned llm. Accessed June 30, 2023. <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.
- Dai, Yaoyao, and Alexander Kustov. 2022. When Do Politicians Use Populist Rhetoric? Populism as a Campaign Gamble. *Political Communication*, <https://doi.org/10.1080/10584609.2022.2025505>. <https://doi.org/10.1080/10584609.2022.2025505>.
- Ganguli, Deep, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, et al. 2022. *Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned*. arXiv: 2209.07858 [cs.CL].
- Hawkins, Kirk A, and Cristóbal Rovira Kaltwasser. 2018. Measuring populist discourse in the united states and beyond. *Nature human behaviour* 2 (4): 241–242.
- Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. *Large language models are zero-shot reasoners*. arXiv: 2205.11916 [cs.CL].
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 7871–7880. Online: Association for Computational Linguistics, July. <https://doi.org/10.18653/v1/2020.acl-main.703>. <https://aclanthology.org/2020.acl-main.703>.
- Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* (New York, NY, USA) 55, no. 9 (January). issn: 0360-0300. <https://doi.org/10.1145/3560815>. <https://doi.org/10.1145/3560815>.
- MosaicML NLP Team. 2023. Introducing mpt-7b: a new standard for open-source, commercially usable llms. Accessed: 2023-03-28. Accessed March 28, 2023. www.mosaicml.com/blog/mpt-7b.
- Mudde, C. 2004. The Populist Zeitgeist. *Government and Opposition* 39 (4): 541–563.
- Mudde, Cas, and Cristóbal Rovira Kaltwasser. 2017. *Populism: a very short introduction*. Oxford University Press.

- Mudde, Cas, and Cristóbal Rovira Kaltwasser. 2018. Studying Populism in Comparative Perspective: Reflections on the Contemporary and Future Research Agenda. *Comparative Political Studies* 51 (13): 1667–1693. issn: 0010-4140. <https://doi.org/10.1177/0010414018789490>. <http://journals.sagepub.com/doi/10.1177/0010414018789490>.
- Müller, Jan-Werner. 2017. *What is populism?* Penguin UK.
- Palmer, Alexis, and Arthur Spirling. 2023. *Large language models can argue in convincing and novel ways about politics: evidence from experiments and human judgement*. Technical report. Working paper), Technical report.
- Penedo, Guilherme, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, arXiv: 2306.01116. <https://arxiv.org/abs/2306.01116>.
- Törnberg, Petter. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.
- Williams, Adina, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, 1112–1122. New Orleans, Louisiana: Association for Computational Linguistics, June. <https://doi.org/10.18653/v1/N18-1101>. <https://aclanthology.org/N18-1101>.
- Wu, Patrick Y, Joshua A Tucker, Jonathan Nagler, and Solomon Messing. 2023. Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. *arXiv preprint arXiv:2303.12057*.
- Wuttke, Alexander, Christian Schimpf, and Harald Schoen. 2020. When the whole is greater than the sum of its parts: on the conceptualization and measurement of populist attitudes and other multidimensional constructs. *American Political Science Review*, 1–19.

Appendix

Table A3. Model performance metrics for measuring populist speeches in the test set.

		Accuracy	Precision	Recall	F-1 Score
ChatGPT 3.5	Populism Long	0.69	0.23	0.55	0.32
	Populism Short	0.60	0.24	0.91	0.38
	Populism (Yes/No)	0.69	0.25	0.64	0.36
	Mudde	0.73	0.26	0.55	0.35
	Aggregated Multi-dimension	0.88	0.6	0.27	0.38
ChatGPT 3.5 Ensemble					
MPT	Populism Long	0.83	0.42	0.72	0.53
	Populism Long	0.49	0.16	0.64	0.25
	Populism Short	0.30	0.11	0.55	0.18
	Populism (Yes/No)	0.54	0.20	0.82	0.33
	Mudde	0.44	0.13	0.55	0.21
Falcon	Aggregated Multi-dimension	0.73	0.08	0.09	0.08
	Populism Long	0.43	0.09	0.36	0.15
	Populism Short	0.71	0.23	0.45	0.30
	Populism (Yes/No)	0.56	0.1	0.27	0.15
	Mudde	0.56	0.04	0.09	0.05
BART	Aggregated Multi-dimension	0.70	0.12	0.18	0.14
	Populism	0.48	0.21	1	0.34
	Three Dimensions	0.40	0.18	0.91	0.29
	Aggregated Multi-dimension	0.59	0.18	0.55	0.27
	Aggregated Multi-dimension Long	0.75	0.29	0.55	0.37
Dai and Kustov (2022) Supervised (test)		0.91	0.45	0.83	0.59