Large Language Models as a Substitute for Human Experts in Annotating Political Text

Michael Heseltine[1] and Bernhard Clemm von Hohenberg[2]

[1]Amsterdam School of Communication Research, University of Amsterdam, m.j.heseltine@uva.nl
[2]GESIS Leibniz Institute for the Social Sciences, Cologne, bernhard.clemm@gesis.org

September 29, 2023

Abstract

Large-scale text analysis has grown rapidly as a method in political science and beyond. To date, text-as-data methods rely on large volumes of human-annotated training examples, which places a premium on researcher resources. However, advances in large language models (LLMs) may make automated annotation increasingly viable. This paper tests the performance of GPT-4 across a range of scenarios relevant for analysis of political text. We compare GPT-4 coding with human expert coding of tweets and news articles across four variables (whether text is political, negativity, sentiment, and ideology) and across four countries (the United States, Chile, Germany, and Italy). GPT-4 coding is highly accurate, especially for shorter texts such as tweets, correctly classifying texts up to 95% of the time. Performance drops for longer news articles, and very slightly for non-English text. We introduce a "hybrid" coding approach, in which disagreements of multiple GPT-4 runs are adjudicated by a human expert, which boosts accuracy. Finally, we explore downstream effects, finding that transformer models trained on hand-coded or GPT-4-coded data yield almost identical outcomes. Our results suggests that LLM-assisted coding is a viable and cost-efficient approach, although consideration should be given to task complexity.

Keywords— large language models; GPT; machine learning; text analysis; text-as-data

Political science has increasingly embraced supervised machine learning as an accurate and cutting-edge tool in the large-scale analysis of political text, greatly supported by ready-to-use methods such as the transformer-based text classifier BERT. Existing applications range from analyses of elite rhetoric (Ballard et al., 2023), to classifications of news sentiment (Rozado et al., 2022), and the detection of hate speech (Mozafari et al., 2020). However, one central limitation of these methods is that each classification requires large amounts of human-annotated training data. Depending on the complexity of the task, reliable modelling requires training data ranging in the 1,000s to 10,000s of annotated text examples, often from multiple coders. This places severe financial and time constraints on researchers.

Recent works have, however, shown the potential for large language models (LLMs) such as the GPT family to perform a range of tasks in the social sciences, including ideological scaling (Wu et al., 2023), the classification of legislation (Nay, 2023), and the detection of hate speech (Huang et al., 2023). LLM classification may therefore be a viable means of reducing manual annotation labour and cutting costs, while providing high levels of classification accuracy or even outperforming human coders (Gilardi et al., 2023; Ornstein et al., 2023; Tornberg, 2023). Additionally, LLMs have also shown the potential for the accurate classification of texts across languages (Kuzman et al., 2023), opening up avenues for research in languages not spoken by researchers.

With these developments in mind, this paper evaluates the potential for large language models to act as a substitute for manual text coding in political science (and potentially beyond). Since previous studies have focused primarily on single tasks (Huang et al., 2023; Kuzman et al., 2023; Tornberg, 2023) or single contexts (Gilardi et al., 2023; Nay, 2023; Ornstein et al., 2023) and do not test downstream effects, further exploration is warranted. We assess the accuracy of coding with GPT-4—the most up-to-date version of the GPT client[1]—across a range of text annotation tasks ubiquitous in political science, namely determining whether text is political, whether it transports negativity (both as a binary and a multi-category variable) and scaling its ideological leaning.

We offer several contributions to the fast-evolving literature on LLM-assisted methodology: First, as the bulk of extant evidence tests GPT coding performance on short text such as tweets

[1]While ChatGPT has been the recent focus of debate, the LLM space is fast-evolving, with ChatGPT (based on GPT v3.5) already superseded by GPT-4.

(but Gilardi et al., 2023), we assess performance also for longer texts, i.e., news articles. Second, few studies test GPT coding performance beyond the U.S., despite its known cultural bias (Johnson et al., 2022). We advance the literature by testing GPT-4 coding accuracy in three other languages and contexts, namely Chile, Germany, and Italy. Third, we complement previous work by testing a "hybrid" coding approach of humans supporting the machine, which is still much cheaper than human coding, but potentially more accurate. Fourth, we explore downstream impacts of differences between expert and GPT-4 coding using a case study of political rhetoric in the U.S. 2022 Congressional primary elections.

The results show, first, that GPT-4 coding can be highly accurate, achieving as much as 91% agreement with expert coding on the classification of political rhetoric, 95% agreement on the classification of negative rhetoric (binary), 82% on sentiment (three categories) and 85% on ideology. Through our hybrid approach, which includes minimal levels of hand-validation (typically less than 10% of the full training set), accuracy of all measures can be further improved. Second, GPT-4's performance drops slightly for full news articles compared to tweets, suggesting potential limitations to automated classification depending on the specific text format. Third, promisingly, GPT-4 generally shows similar (though slightly lower) levels of accuracy in non-English classification of tweets, suggesting that GPT-4 coding is a viable option for researchers working with data across languages. Last, downstream, the modeling based on manual and GPT-4 coding produced almost identical results, suggesting that the level of disagreement between human and GPT-4 coding may have minimal implications for substantive research. We consider the limitations of our approach in more depth in the Discussion.

Data and Method

Our baseline test of accuracy is based on a sample of 635 tweets from Members of Congress in the United States, randomly chosen out of all their tweets from between 2009 and 2022. To test the effect of text length on classification accuracy, we also collected a random sample of 200 news articles from 2016 and 2017 across a range of U.S. news outlets (NYT, WaPo, Bloomberg, Breitbart,

Vox, The Atlantic). To test accuracy across languages and contexts, we further selected a random sample of tweets from all tweets posted by members of parliament between 2009 and 2022 in Chile (330 tweets), Germany (700 tweets) and Italy (330 tweets).[2] Although this selection of countries is by no means exhaustive and was influenced by our own expertise and access to expert coders, this multi-country approach still goes beyond existing U.S.-focused evidence.

Manual expert coding. Experts coded the sets of tweets (U.S., Chile, Germany, Italy) across four dimensions, according to detailed instructions shown in SI C: (1) whether the text was political or not (binary); (2) whether the text contained negative messaging or not (binary); (3) whether the text contained negative, positive, or neutral messaging (three categories); (4) whether the text was ideologically left-wing, centrist, or right-wing (three categories). The U.S. data were coded by two coders, with any discrepancies then mutually resolved, with the final coding serving as "ground truth" for our accuracy assessments. Non-U.S. tweets were single-coded by an expert of the respective country, with a "ground truth" review then conducted based on translation and confirmation with a second reviewer. For the test of varying text length, U.S. news articles were coded for only the binary political and negativity criteria.[3]

GPT-4 coding and performance assessment. For each of the four coding tasks, we prompted GPT-4 twice with coding instructions aligning with those given to the human experts (see SI C). For each concept, therefore, the data have scores from two GPT-4 "coders", which we refer to as "GPT-4 first run" and "GPT-4 second run". In an alternative approach, we also try a simple prompt that just mentions the concept and gives no further explanations. Full comparisons with this "zero-shot" approach are shown in SI E. Due to message length restrictions, tweets were classified in batches of twenty, with each batch run in a fresh instance of the GPT-4 chat client to avoid any biasing from previous prompts. News articles were classified in batches of four. To quantify the accuracy of GPT-4 coding, we present the percentage of classifications in each run which agree with the final expert coding (alternatively, results using F1 scores are also presented in

[2]Some non-U.S. tweets were actually written in English, but these were left in as a test of how GPT-4 handled the annotation of multiple languages within a single batch. There appeared to be no issues.

[3]News articles were deemed to be too rarely "positive" for a multi-category sentiment classification and the ideological classification of primarily fact-based reporting was deemed to be unfeasible.

SI D). For all four concepts, we start with a "baseline accuracy" for U.S. tweets, before we moving on to news articles, and then non-U.S. tweets.

Hybrid Human-GPT-4 coding. We further exploit the fact that the two rounds of GPT-4 coding based on the exactly the same instructions and data will, due to randomness, yield slightly different results. This disagreement likely occurs on edge cases, which represent important nuance in any given concept. We therefore test a "hybrid" model, where disagreements between the "GPT-4 first run" and the "GPT-4 second run" are adjudicated by a human expert. Of course, this adjudication process pushes the classification more towards the expert coding (although only for contested cases) and is therefore likely to improve the accuracy. However, the results illustrate that LLM-assisted coding can be optimized through minimal additional human effort. In the results, we also report the frequency of disagreement between the two GPT-4 runs as an indicator of additionally required human input.

Downstream effects. Finally, we expert-coded, GPT-4-coded (twice), and hybrid-coded a set of 3,000 additional tweets from candidates in the 2022 U.S. congressional elections for both negative messaging (binary concept) and political ideology. Based on these four codings, we fine-tuned four models of negative messaging and four models of ideology using BERTweet (Nguyen et al., 2020), a transformer package designed specifically for handling social media data. We use the trained classifiers to predict negativity and ideology in all tweets (excluding retweets) sent by congressional candidates before their state primary in 2022. The resulting classifications are then compared side-by-side in both descriptive and predictive models to test for meaningful differences in the resulting analyses.

Classification Performance

Baseline accuracy

Beginning with the U.S. tweets, Figure 1 below shows the percentage agreement between the expert coding and the two GPT-4 coding runs. We report F1 scores as an alternative measure in SI D, with substantively identical results. Bars are color-coded by classification approach, indicating whether the result is based purely on GPT-4 coding or also includes an expert reconciliation of disagreements between GPT-4 runs (i.e., the hybrid approach). Note that all results are based on GPT-4 prompts using full coding instructions. In SI E, we also report results when prompting GPT-4 just with the concept of interest without defining the concept further. In most cases, including details improves accuracy.

The results, overall, show a relatively high degree of accuracy, but also highlight some important variance across classification tasks. Beginning with political classification, the two rounds of GPT-4 coding agreed with expert coding 88.3% and 91.1% of the time. When reconciling disagreements between GPT-4 rounds using human validation (7.6% of instances), this accuracy increases to 93.4%. For the binary negative classification task, accuracy is even higher, with the two coding rounds agreeing with the expert coding 94.5% and 94.3% of the time. The hybrid validation approach improves these results further to 96.9% agreement, based on 4.9% disagreement between GPT-4 rounds. In general, for these two binary classification tasks, GPT-4 results closely approximate human annotation.

Beyond the binary tasks, however, accuracy does drop. In the case of three-category sentiment coding, this decrease is particularly notable. The GPT-4 coding rounds were accurate 81.7% and 80.6% of time, with the hybrid approach then increasing this accuracy to a respectable 86.6% (based on 13.4% GPT-4 disagreement). For context, the rate of agreement between human coders on this concept was 87.2%, suggesting that GPT-4 does underperform human coding accuracy, but not to an extreme degree. For ideology, the level of accuracy (84.7% and 85%) is lower than in the binary tasks. With the hybrid approach, based on 10.6% disagreement between rounds, accuracy improves to over 90%. For reference, the baseline rate of agreement between human coders was

also 85%. Collectively, given the complexity of the task at hand, the results actually highlight the likely strength of GPT-4 in this particular coding task.

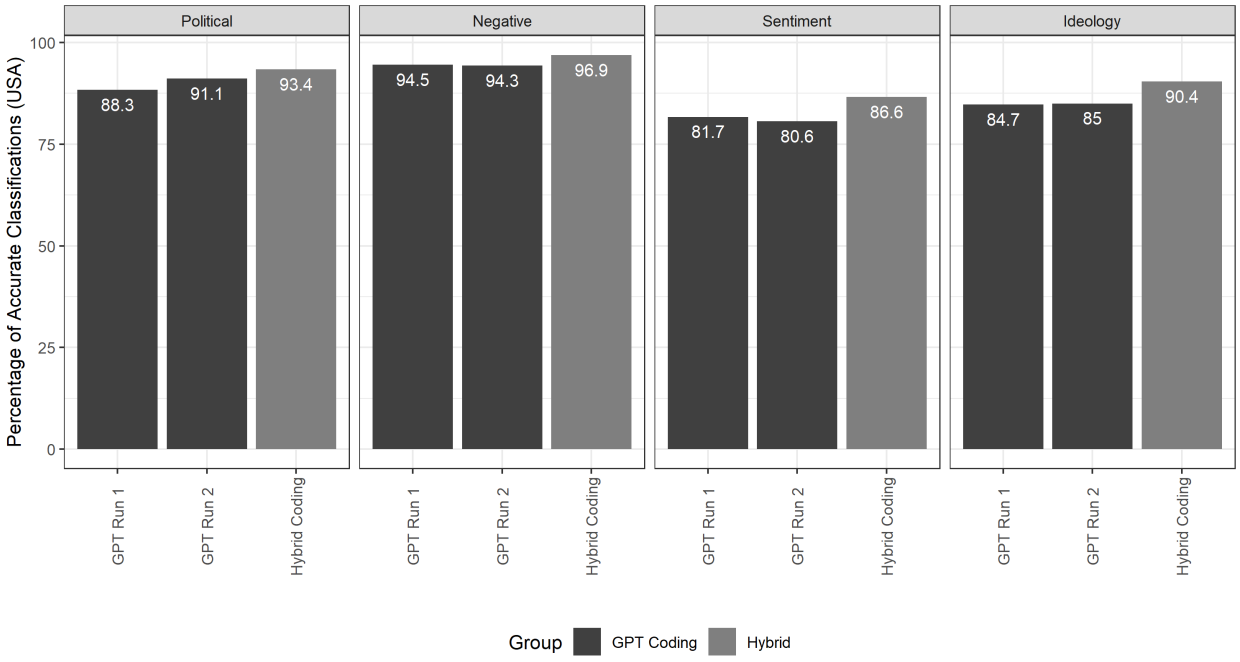

Figure 1: Classification accuracy rates for U.S. tweets, by task and coding method.

Accuracy for longer texts

Classification accuracy may change when applied to differing text types, especially in terms of the overall length of the text. Indeed, when classifying full news articles, some notable changes in accuracy occur. As illustrated by Figure 2, for the classification of political, accuracy increases slightly to 96.5% and 95%, while the accuracy of negativity classification drops dramatically to 80% and 76%. Evaluating the divergences qualitatively, longer texts appear to be providing differing cues for the two types of classification task. For political classification, longer text provides greater context and more opportunities for political keywords. For negativity classifications, however, longer text provides more conflicting signals, with single articles often containing positive, neutral, and negative components. However, given the black-boxiness of LLMs, we ultimately do not know what the reason for the decrease in performance is. In any case, researchers should consider the combination of text type and classification task when deciding about the viability of GPT-4 for

their coding requirements.

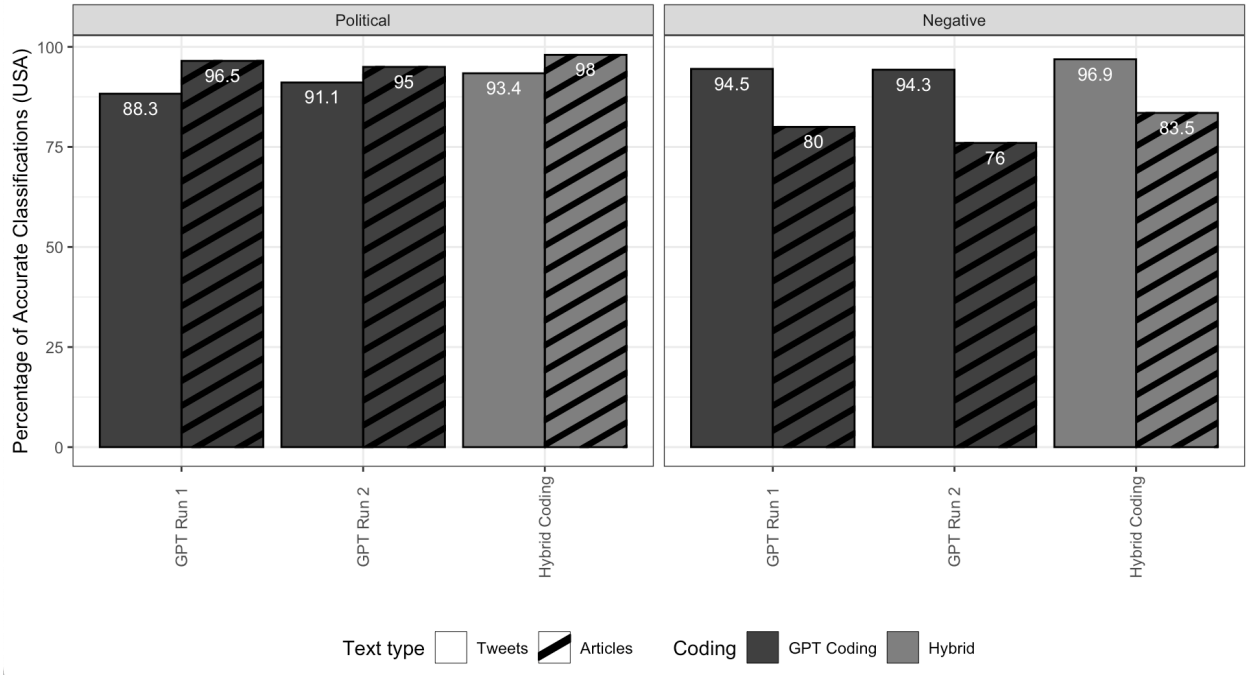

Figure 2: Classification accuracy for U.S. tweets and news articles, by task and coding method

Accuracy across languages

Having established performance levels on English-language data, the question is whether GPT-4 will perform consistently in other languages. Figure 3 shows the percentage agreement of GPT-4 with expert coding of tweets by Italian, German and Chilean politicians. Performance is very similar across languages. When classifying political tweets, accuracy is above or just below 90% across countries and runs, closely tracking accuracy rates in the U.S. context. In terms of negativity classification, accuracy is still high, but somewhat lower in Germany and Italy, sitting just above 85% as opposed to above 90% in the U.S. Accuracy for the three-way sentiment classification again drops to below 80% in all countries, a potentially unsatisfactory result. For the ideology classification, results are strong (between 81% and 84% across countries), but again, just below the level of accuracy seen in the U.S. Again, the hybrid coding approach increases overall accuracy, bringing accuracy in many contexts above or approaching 90%.

Based on these results, the potential for simultaneous translation *and* classification of text is a particularly appealing opportunity for automated coding approaches. In some tasks, performance is marginally lower than in the U.S., although still strong, while in others (especially the classification of political content) results are largely indistinguishable from the classifications in the U.S. context.

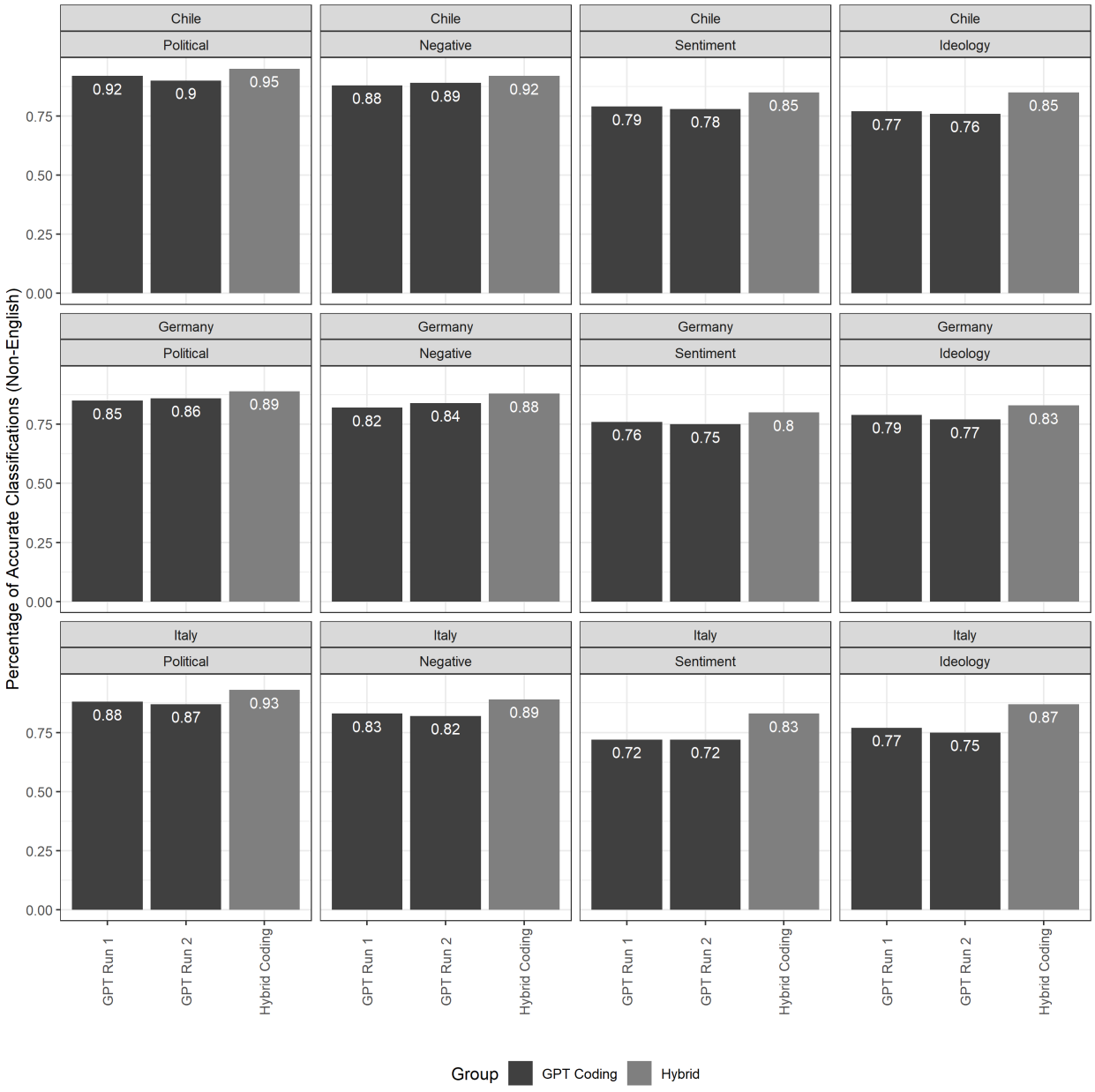

Figure 3: Classification accuracy for tweets from Chile, Germany and Italy, by task and coding method.

Downstream Effects: Congressional Primary Case Study

Although the differences in annotation results may be minimal between GPT-4 and expert coders, they may still have significant downstream impacts on modelling and results, especially if GPT-4 coding is systemically biased. Therefore, to assess whether the differences are meaningful, we offer two insights from a U.S. case study, based on expert-coded and GPT-4-coded versions of the binary negativity classification, as well as the three-way ideology classification discussed above.

Our case study connects to long-running threads of research about negativity and ideological messaging during political campaigns. Studies have found that negativity fluctuates across the course of a campaign (Lau & Rovner, 2009), with negative messaging often increasing as the general election approaches, (Hassell, 2021), while, conversely, decreasing prior to primary elections (Peterson & Djupe, 2005). Similarly, research has shown that candidates may be incentivized to vary their publicly presented ideology across different stages of a campaign (Brady et al., 2007).

For our case study, we trained a total of eight NLP models (four for each of the two variables of interest) based on the different classification approaches presented above. We use these models to test whether levels of negativity and the ideology in Congressional candidate messaging changes in the run-up to the U.S. 2022 Congressional primary date, using tweets sent within the final 90 days of each campaign. To do this, we split a set of 3,000 tweets (distinct from the set discussed above) into 2,500 training and 500 test examples. We coded these for negativity and ideology, first, by hand, and second, with two GPT-4 runs. Where disagreement occurred between GPT-4 coding runs, an expert coder adjudicated the disagreement to create the fourth hybrid coding. We used the 600 tweets discussed above as a validation set. These data were then used to train four separate BERTweet (Nguyen et al., 2020) models for each classification task (GPT-4 Run 1; GPT-4 Run 2; manual coding; hybrid coding), each of which then predicted negativity and ideology of all tweets sent by congressional candidates in the U.S. House and Senate primaries in 2022. The descriptives, trends, and modelling we present below are based on 391,973 tweets in total.

Figure 4 below shows the daily percentages of tweets classified as negative in all four classifications side-by-side. Importantly, average negativity across the period are similar across all four

methods: The manual coding classifies 31.1% of tweets as negative, the two pure GPT-4 runs classify 29.5% and 29.3% as negative, respectively, and the hybrid approach classifies 30.2% of tweets as negative. Additionally, the over-time trends are almost identical across all four models, with negativity being relatively steady in the run-up to the election with a perceptible slight decrease in the week before the election.

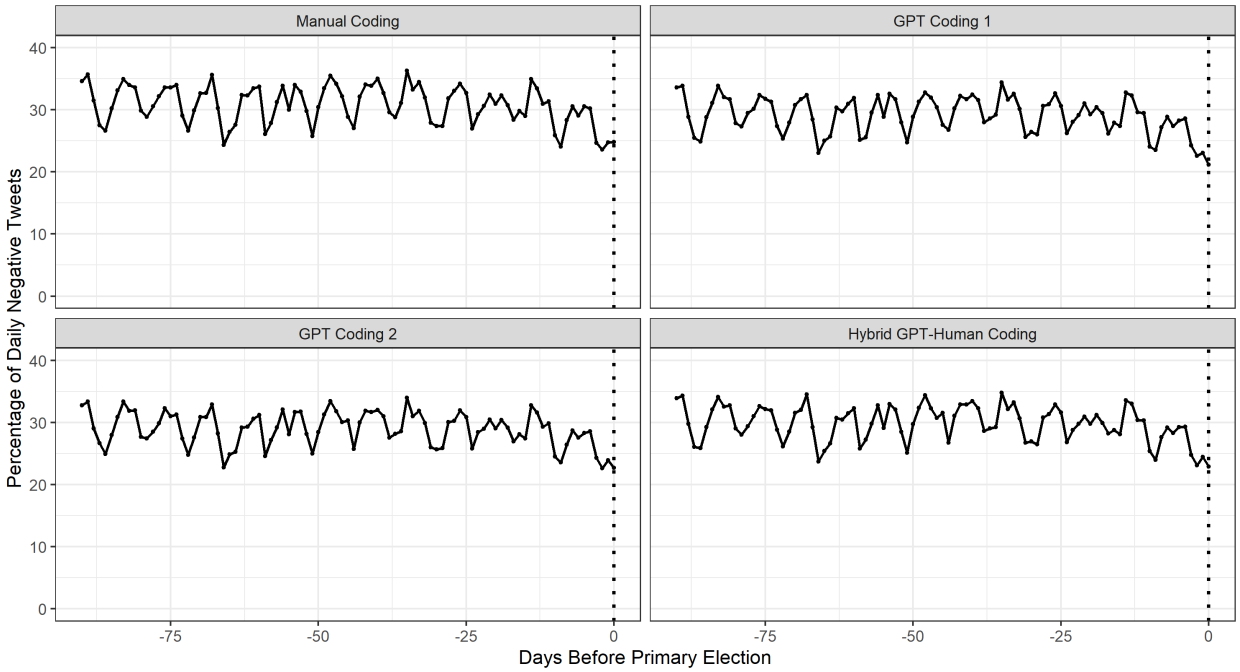

Figure 4: Daily Trends in Tweet Negativity Based on the Four Different Classifiers.

Figure 5 shows the daily trends in tweet ideology, aggregated by political party, with results closer to -1 indicating more ideologically liberal content. Again, the results are almost identical across the four methods. The overall Democratic and Republican party averages, respectively, are 0.79 and 1.24 in the manual coding model, 0.75 and 1.19 in the first GPT-4 model, 0.82 and 1.26 in the second GPT-4 model, and 0.81 and 1.25 in the hybrid model. Over time, all models show identical trends, with no signs of any moderation or increased extremity directly before the primary date.

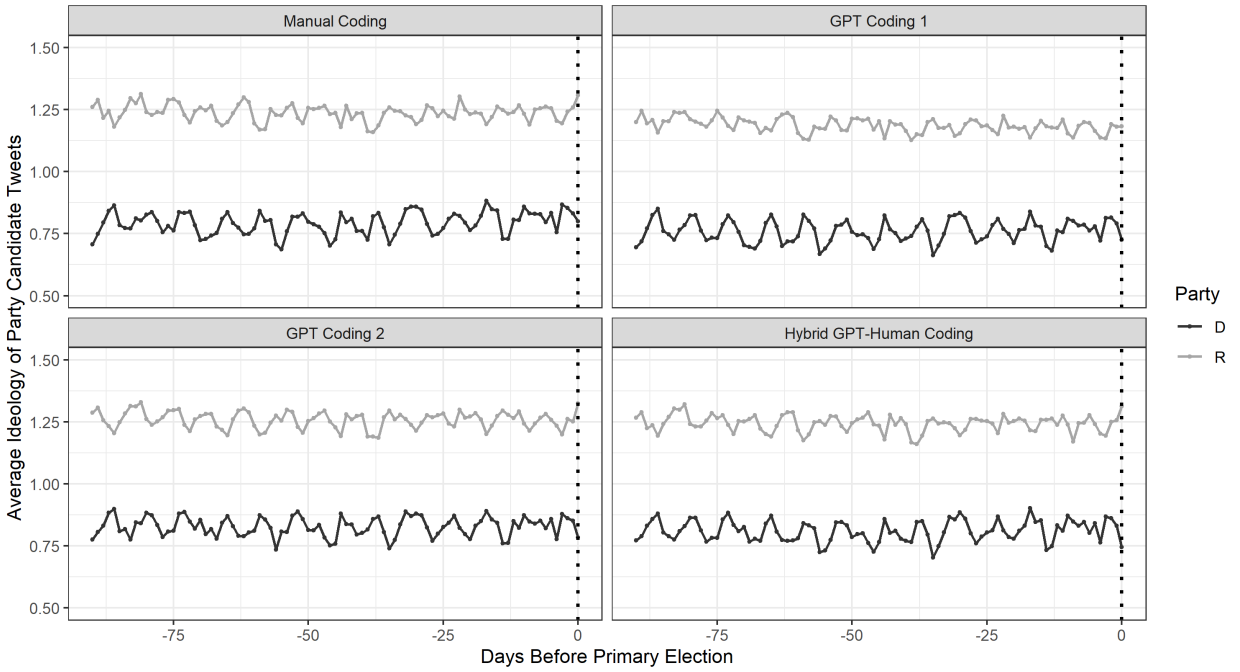

Figure 5: Daily Trends in Tweet Ideology, by Party, Based on the Four Different Classifiers.

Thus, the two over-time plots provide consistent evidence that across both measures of interest, central descriptive findings are identical independent of whether the training data were manually coded, GPT-4-coded, or coded using our hybrid approach.

Moving beyond descriptive trends, we test the consistency of the methods when modelling the two concepts of interest as a dependent variable. Figures 6 and 7 below show, at the candidate level, linear regressions predicting the percentage of negativity and the average ideology of messages sent by a candidate in the pre-primary period, based on a set of key covariates (see SI F for details). As can be seen, results across models based on the four different codings are almost identical, with no changes in significance of any predictors across models. As such, using either hand-coded, GPT-4-coded, or hybrid-coded training data to explore candidate messaging in the 2022 congressional primaries produces generally indistinguishable results.

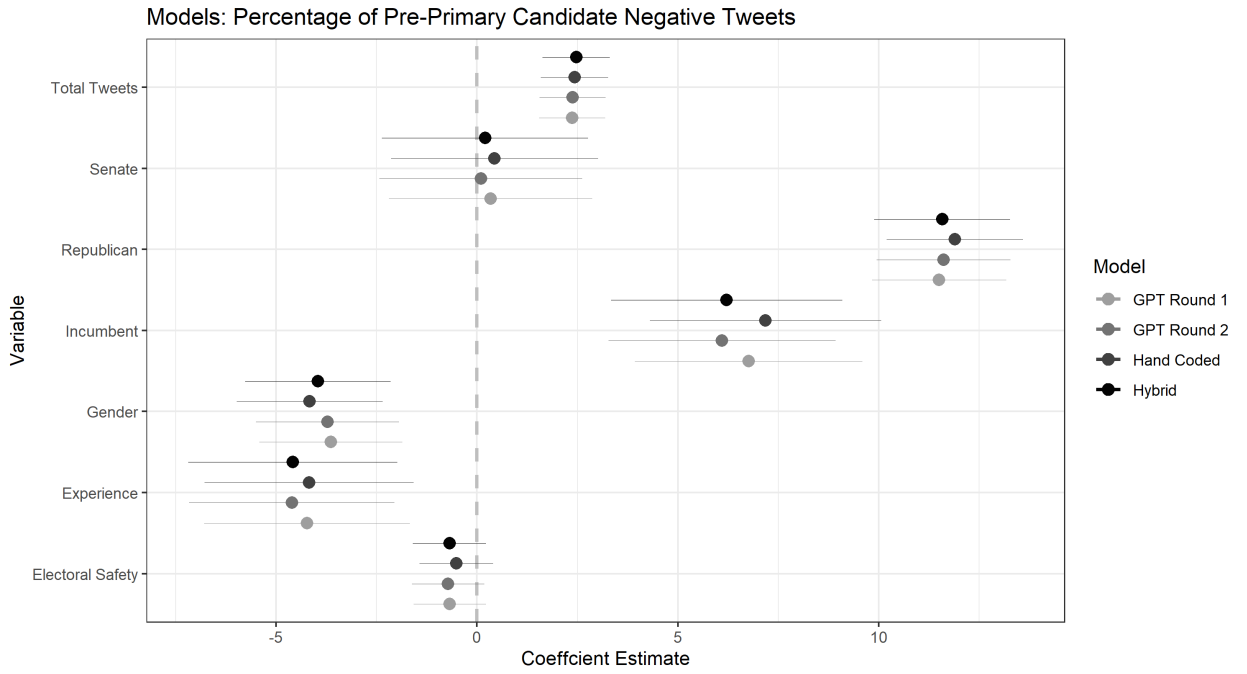

Figure 6: Coefficient Estimates for Predictors of Percentage of Negative Twitter Messaging from a Given candidate.

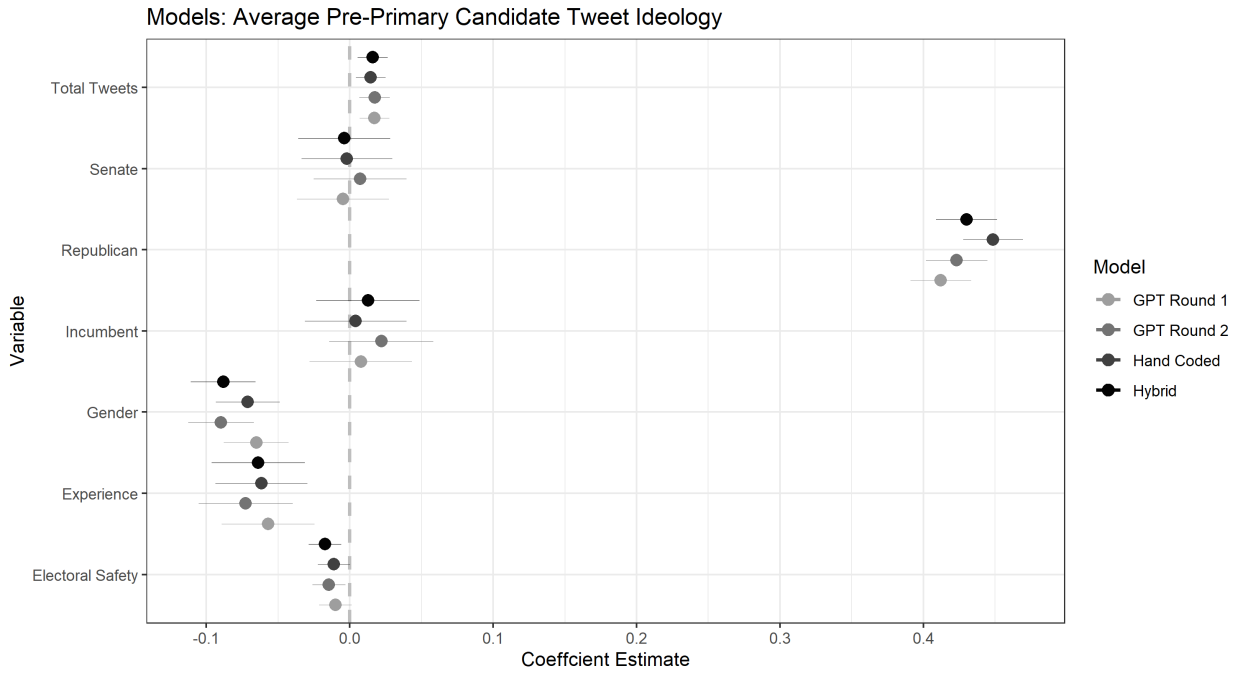

Figure 7: Coefficient Estimates for Predictors of the Average Ideology of Twitter Messaging from a Given candidate.

Discussion and Conclusion

In this study, we assessed the potential of GPT-4 (or similar LLMs) to accurately substitute for manual human text annotation, with a particular focus on applications in political science. The results show that GPT-4 coding, benchmarked against the "ground truth" of expert coding, is highly accurate and demonstrates clear potential for use across a range of research scenarios. The edge of our hybrid coding approach over pure GPT-4 coding suggests the following three-stage process for researchers interested in LLM-supported coding: First, classify all training text at least twice using a LLM. Second, manually reconcile discrepancies between the two rounds of coding. Third, use this single reconciled training set for downstream tasks such as training transformer models. While this approach will not perfectly replicate an expert manual coding approach on all tasks, our results indicate that differences in downstream applications may be negligible.

One might also wonder why it is necessary to limit ourselves to using LLMs to merely classify training data instead of all data, given the accuracy. Indeed, side-by-side comparisons suggest that ChatGPT and BERT models can produce similar classification results (Zhong et al., 2023). However, at present, the rate of classification through GPT-4 is considerably slower than, for example, a transformer model classifying data on a high-end GPU. Non-English GPT-4 coding in our project was even slower, given the integrated language detection and translation inherent in the classification process.

Despite the promising results, our approach has important limitations. First, researchers should consider the complexity of the classification objects, as GPT-4 performed worse on longer, more complex texts. Second, it is also unclear how GPT-4 would perform for even more complex concepts. We noticed a performance drop for the more complex three-category classifications, compared to the binary concepts. However, we note that simple classifications such as whether content is "political" or "negative" are very common in the field, and even human experts may not always disagree on how to code text on more complex dimensions. Third, we cannot say how LLM-assisted coding would do in other, non-Western contexts. We already noticed a small drop in performance for non-English texts, although results looked still satisfactory.

Finally, standard limitations in the everyday use of LLMs also apply to their usage for classification tasks. Biases inherent in the training of these models (Bisbee et al., 2023; Motoki et al., 2023) may seep into text annotation, especially ones more specific or contentious than classifications done here. Researchers should be mindful of these potential biases and carefully consider their impact on potential outcomes.

The implications of our findings are potentially substantial. All GPT-4 coding for this project was completed using a single $20 monthly subscription. Hence, financial disparities between researchers effectively evaporate for tasks where LLMs can substitute human labour. By achieving comparable results at a fraction of the time and cost, GPT-4 coding opens up machine learning applications to an incredibly diverse pool of researchers, benefiting the discipline through new perspectives, datasets, and areas of focus. High levels of cross-language accuracy provide significant opportunity and incentive for researchers to increase the levels of comparative studies. The field of political science may benefit from more generalizable, global work, with less of a targeted focus on single regions and countries (especially the United States).

References

Ballard, A. O., DeTamble, R., Dorsey, S., Heseltine, M., & Johnson, M. (2023). Dynamics of polarizing rhetoric in congressional tweets. *Legislative Studies Quarterly*, *48*(1), 105–144.

Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., & Larson, J. (2023). Artificially precise extremism: How internet-trained llms exaggerate our differences. *SocArXiv*. osf.io/preprints/socarxiv/5ecfa

Brady, D. W., Han, H., & Pope, J. C. (2007). Primary elections and candidate ideology: Out of step with the primary electorate? *Legislative Studies Quarterly*, *32*(1), 79–105.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv*. https://doi.org/10.48550/arXiv.2303.15056

Hassell, J., Hans. (2021). Desperate times call for desperate measures: Electoral competitiveness, poll position, and campaign negativity. *Political Behavior*, *43*, 1137–1159.

Huang, F., Kwak, H., & An, J. (2023). Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv*. https://doi.org/10.48550/arXiv.2302.07736

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022). The ghost in the machine has an american accent: Value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Kuzman, T., Mozetic, I., & Ljubešic, N. (2023). Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification. *ArXiv, abs/2303.03953*.

Lau, R. R., & Rovner, I. B. (2009). Negative campaigning. *Annual Review of Political Science*, *1*, 285–306.

Motoki, F., Neto, V. P., & Rodrigues, V. (2023). More human than human: Measuring chatgpt political bias. *Public Choice*, 1–21.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on bert model. *PLoS ONE*, *8*.

Nay, J., John. (2023). Large language models as corporate lobbyists. *arXiv*. https://doi.org/10.48550/arXiv.2301.01181

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14.

Ornstein, J. T., Blasingame, E. N., & Truscott, J. S. (2023). How to train your stochastic parrot: Large language models for political texts. *Working Paper*. https://joeornstein.github.io/publications/ornstein-blasingame-truscott.pdf

Peterson, D. A. M., & Djupe, P. A. (2005). When primary campaigns go negative: The determinants of campaign negativity. *Political Research Quarterly*, *58*(1), 45–54.

Rozado, D., Hughes, R., & Halberstadt, J. (2022). Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models. *PLoS ONE*, *10*.

Tornberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv*. https://doi.org/10.48550/arXiv.2304.06588

Wu, P. Y., Tucker, J. A., Nagler, J., & Messing, S. (2023). Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. *arXiv*. https://doi.org/10.48550/arXiv.2303.12057

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv*. https://doi.org/10.48550/arXiv.2302.10198

Supplementary Information

A Coding Breakdown

Three individual coders were collectively responsible for all hand-coding used in this analysis. The first coder, the corresponding author, coded all English language text across all classification tasks. The second coder, an assistant fluent in English, Spanish and Italian with expertise in the respective political contexts of the U.S., Chile and Italy, coded all Spanish and Italian text as well as acted as a second coder on some English text. A third coder, the second author, expert in U.S. and German politics, coded all German text and additional English language text. A complete breakdown of coding combinations on each text type and classification type is shown below.

Text Type	**Classification Task**	**Coder1**	**Coder2**	**Coder3**	**Coder Reconciliation**	**GPT-4 Reconciliation**
U.S. Tweets	Political	100%	100%	NA	Disagreements mutually reconciled by both coders	Reviewed by coder 2
U.S. Tweets	Negative	100%	100%	NA	Disagreements mutually reconciled by both coders	Reviewed by coder 2
U.S. Tweets	Sentiment (Multi)	100%	NA	100%	Disagreements mutually reconciled by both coders	Reviewed by coder 2
U.S. Tweets	Ideology	100%	NA	100%	Disagreements mutually reconciled by both coders	Reviewed by coder 3
Media Articles	Political	100%	100%	NA	Disagreements mutually reconciled by both coders	Reviewed by coder 1
Media Articles	Negative	100%	100%	NA	Disagreements mutually reconciled by both coders	Reviewed by coder 1
Spanish Tweets	All	NA	100%	NA	NA	Tweets translated by coder 2 for review by coder 1
Italian Tweets	All	NA	100%	NA	NA	Tweets translated by coder 2 for review by coder 1
German Tweets	All	NA	NA	100%	NA	Tweets translated by coder 2 for review by coder 1
U.S. Tweets (Full Training Data)	Negative	100%	NA	NA	NA	Reviewed by coder 2
U.S. Tweets (Full Training Data)	Ideology	100%	NA	NA	NA	Reviewed by coder 2

B Coding Instructions

B.1 Political:

In general, references to political developments, political actors, or political topics (abortion, the economy, gun control, environmental regulation etc.) are all included, be it at the international, national, or local level. References to federal organisations are political, as are references to branches of government. Importantly, references to financial institutions and national economic developments are political (as in the fed raises interest rates, or GDP estimated to increase by 1% this year), but references to individual stock prices or company performances are not. We very much want to avoid including content like "Tesla stock falls 10% in June" or "xyz becomes CEO at Blackrock" from being classed as political. Additionally, crime is not always political. "man arrested for drunk driving" is not political. Nor is "Sherriff arrests 3 in shoot out". Crimes with a racial element or that have some connection to politics, like anti-immigrant crimes, are included.

B.2 Negative:

This is specifically "negative messaging". This definition is designed to be distinct from naive definitions which do not separate negative sentiment from merely negative events or outcomes. This therefore includes direct attacks, criticism, disparaging comments, or generally comments which expressly say or at least suggest that some person/event/outcome is bad/negative. For example, while "Hurricane kills 50" is certainly bad, it is not negative messaging. Alternatively, "Failure to address climate change means 50 have died in hurricane" is negative. As such, negative is defined as a message which expresses negative sentiment directly towards an individual, group, or outcome and is not merely any message which contains negative valence.

B.3 Sentiment (3 Category):

As with the binary classification of negativity, positive is defined as a message which expresses positive sentiment directly towards an individual, group, or outcome and is not merely a message which mentions something positive. Therefore, "man wins lottery" is not positive but "really happy to see that my neighbour won the lottery" is. In a political context, this means that matter of fact statements about events such as bills passing or being introduced are taken as neutral. Sometimes messages contain multiple elements, but intent is again important. For example, "I introduced a bill to end childhood hunger because the government has failed our children" is negative.

B.4 Ideology:

This may vary depending on country context, but there are three basic codes: 0 = ideologically left, 1 = ideologically neutral or non-ideological, 2 = ideologically right. First, tweets with no ideological content are classified as neutral. Centrally, two types of content are generally included as ideological. The first is content that would be typically associated with one side or the other. In the US, this means that things like climate change awareness messages are classed as 0 and messages rejecting climate change are a 2. Additionally, messages which attack the other side are also associated with the respective party ideologies. "The Republican party doesn't care about childhood hunger" would be a 0, therefore. Similarly, claiming ownership for one side counts, so "As Democrats, we are the only party working to end childhood hunger" is a 0, even if a general mention of childhood hunger by itself would probably not be classified as either a 0 or a 2. For our purposes, content that is not political like a happy birthday message is classified as a 1, along with all content that does not necessarily have an ideological slant to it or isn't associated with one side or the other. Vague general messaging, even of topic areas is a 1. When there is a clear slant to a policy discussion, this is assigned to the relevant ideology, so "we should improve healthcare" is neutral, while "we need better healthcare by making it free for everyone" is a 0 and "we can improve healthcare by increasing private market options" is a 2.

C Classification Prompts

Below are the various prompts fed into the GPT-4 interface to produce the coding results for each type of classification approach. In this case, results were printed through the widely accessible chat interface. If using the API to create classifications of data, wording would of course need to be modified slightly, asking for an appended classification output, for example, rather than asking for the creation of a new table. For more complex definitions, or to hone precision of concept even further, researchers could of course provide even more detailed definition and coding instructions, similar to the types of instructions given to a human coder.

C.1 Political (no instructions):

create a table with a classification of the following messages as political (1) or non-political (0).

C.2 Negative (no instructions):

create a table classifying the following messages as negative (1) or not negative (0):

C.3 Political (with instructions):

create a table with a classification of the following messages as political (1) or non-political (0). Political is defined as any message which is directly about a political topic, references political developments, or makes reference to a political figure, group, or agency. References to federal organisations are political, as are references to branches of government. Broad mentions of national economic developments are political, but discussions of individual stock prices are not:

C.4 Negative (with instructions):

create a table with a classification of the following messages as negative (1) or not negative (0). Negative is defined as a message which expresses negative sentiment directly towards an individual, group, or outcome and is not merely any message which contains negative valence:

C.5 Negative/Neutral/Positive (with instructions):

create a table with a classification of the following messages as negative (0), neutral (1), or positive (2). Negative is defined as a message which expresses negative sentiment directly towards an individual, group, or outcome and is not merely any message which contains negative valence. Positive is defined as a message which expresses positive sentiment directly towards an individual, group, or outcome and is not merely a message which mentions something positive. Matter of fact statements about events such as bills passing or being introduced are taken as neutral:

C.6 Ideology, Validation Set:

Create a table classifying the following tweets as ideologically liberal (0), ideologically neutral (1), or ideologically conservative (2). Ideology here is defined in the context of the [United States/German/Italian] political system. Tweets with no ideological content are classified as neutral:

C.7 Ideology, Full Models Training Data:

Create a table classifying the following tweets as ideologically liberal (0), ideologically neutral (1), or ideologically conservative (2). Ideology here is defined in the context of the United States political system during the Biden presidency. Tweets with no ideological content are classified as neutral:

D Alternate Accuracy Measure: F1 scores

While the raw percentage accuracy rates of each coding round shown in the main paper are revealing, it is also possible that the high frequency of political messages and low frequency of negative messages produces somewhat misleading accuracy results (classifying everything as 0 for negativity, for example, would still produce as a high accuracy score). Addressing this, Figures A1 through A3 below show the classification results using macro F1 scores, a measure which accounts for the frequency of each category in producing an accuracy score. Using this alternative measure, the central trends found in the main paper hold. F1 scores, are again, very high and show expected changes between classification approaches. The same is true of F1 scores for the international classifications, which are consistent with U.S. message scores.

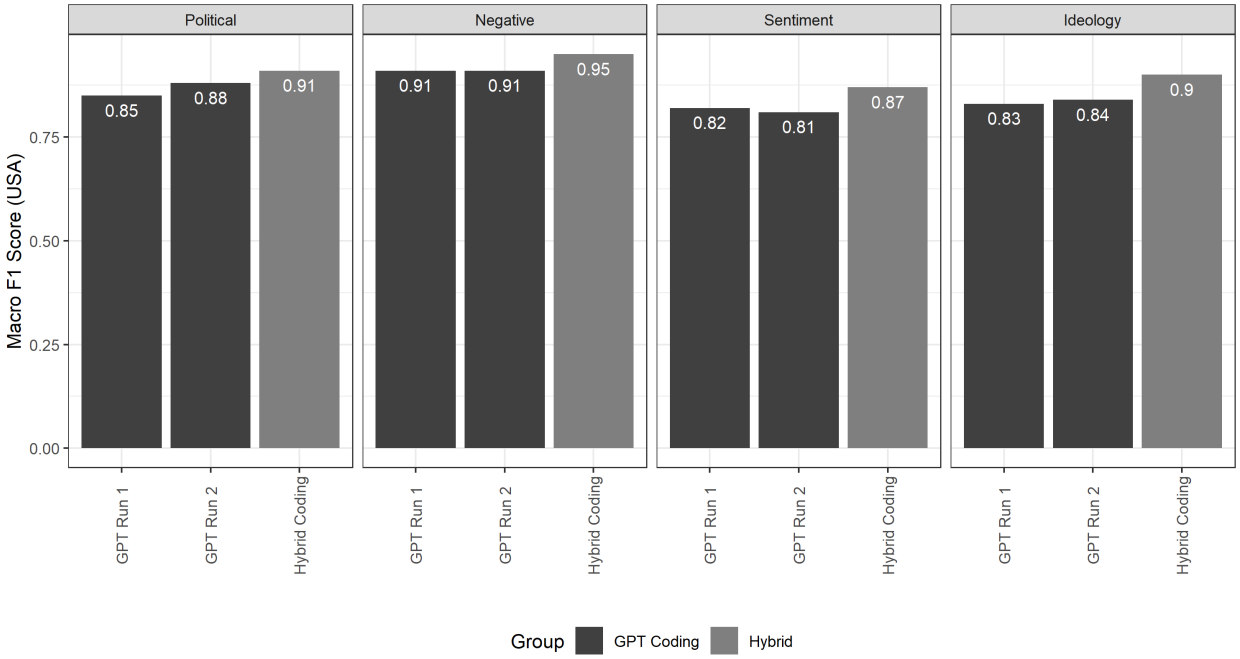

Figure A.8: Classification F1 scores for U.S. tweets, by task and coding method.

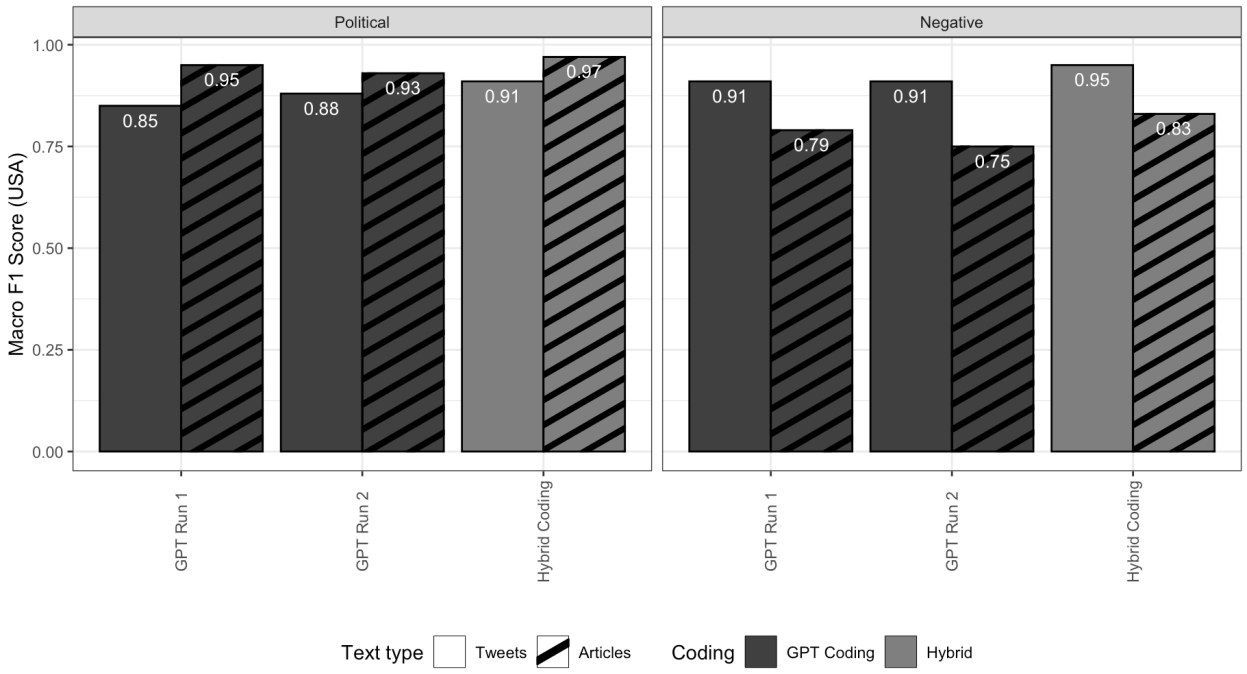

Figure A.9: Classification F1 scores for U.S. tweets and news articles, by task and coding method

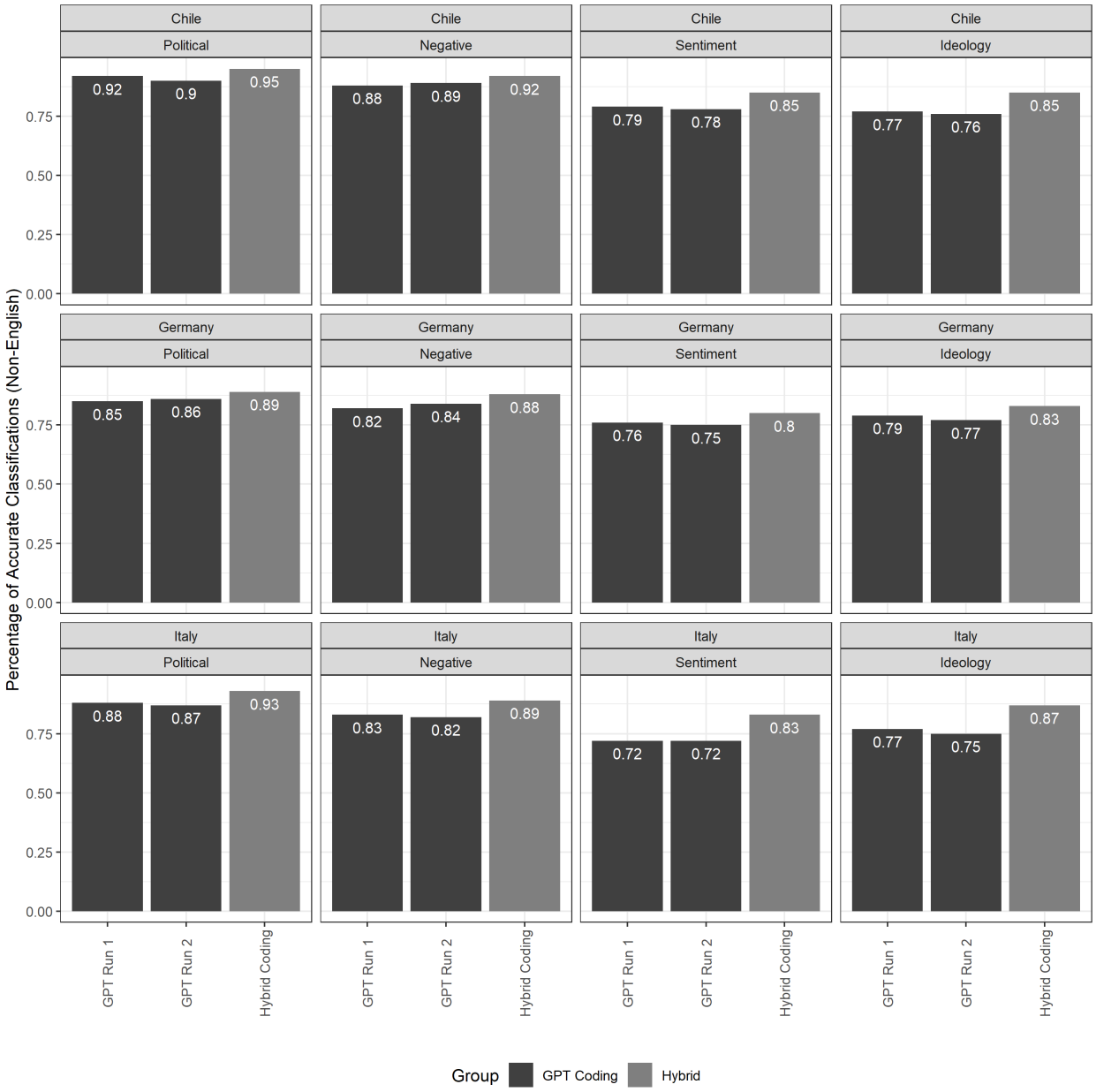

Figure A.10: Classification F1 scores for tweets from Chile, Germany and Italy, by task and coding method.

E Comparison of Coding Performance With and Without Definition Instructions

When no coding instructions were provided, the accuracy (the rate of agreement between the expert coding and the GPT-4 coding) of the political classification was 89% and 88% in the two coding rounds, matching the performance of the classification rounds when instructions

were included. The accuracy of the negative classifications was 90.2% in both rounds, around four percentage points lower than when coding instructions were included. For the three-way sentiment classification, the difference is more significant, with both rounds of coding with no instructions accurately classifying tweets around 70% of the time, while the addition of coding instructions increased this level of accuracy by just over 10 percentage points in each round to over 80%.

These results suggests that, at a basic level, GPT-4 understands the concepts of "political" and "negative" in a very similar way to how social science researchers would define the concept, meaning that GPT-4 can likely be used to identify these concepts with a very high degree of accuracy in an "off the shelf" or zero-shot fashion. For the more complex sentiment task however, over-classification of positive messages was common and was only remedied (in-part) by the addition of instructions. This does suggest that, in the same way that human coders are often trained and provided with definitions and coding instructions, additional detail in prompts can improve GPT-4 classification accuracy of certain concepts. It is therefore recommended that researchers apply at least basic instructions to their prompts to ensure higher levels of coding accuracy.

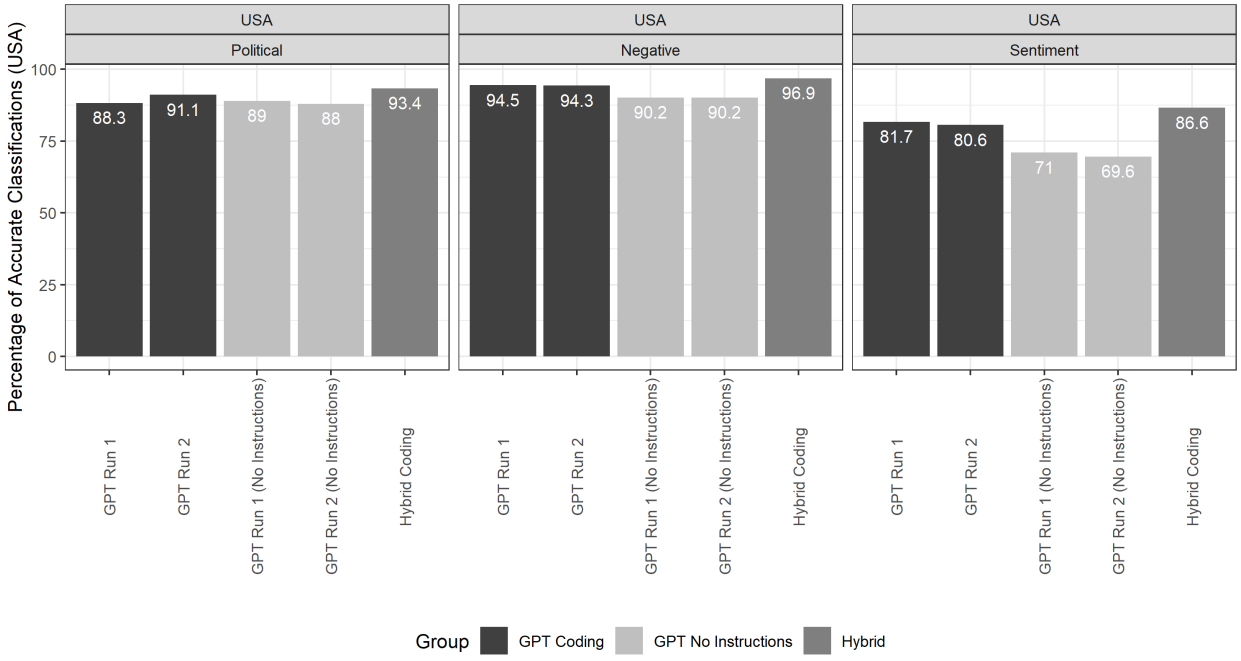

Figure A.11: Classification accuracy with and without coding instructions included in GPT-4 prompts

F 2022 Case Study Data, Covariates, and Results

As this research note focuses primarily on the methodological aspect of GPT-4 classification comparisons, information on the data collection and construction of covariates for the

regression modelling in the 2022 primary case study were not outlined in full in the main paper. Instead, information on these is outlined here below:

F.1 Data

Information on all major party candidates running in a House or Senate primary was hand-collected by the author, including personal, electoral, and social media information. In total, 1,814 candidates had one or more Twitter handles and sent at least one tweet between January 2021 and their 2022 primary election date. 1,719 of these candidates sent tweets within the chosen 90 day pre-primary window. Within this 90 day window, the median candidate sent 154 tweets and 35 negative tweets (as classified by the hand-coded model).

F.2 Variables

Chamber: Binary for whether the candidate is running for the House or Senate

Incumbent: Binary variable for whether the candidate is the incumbent in the district/state.

Experience: Binary variable reflecting whether the candidate has ever previously held elected office at any level (all incumbents are 1 by definition).

Partisan Electoral Advantage: Continuous variable taken as the Cook PVI (Partisan Voter Index) score for a district/state relative to the candidate's party.

Gender: Binary variable for whether the candidate identifies as male or female (one non-binary candidate was excluded for simplicity of modelling).

F.3 Regression Results

Table A.1

	Dependent variable: Pct. Negative Messages			
	Negative Hand	Negative Code1	Negative Code2	Negative Hybrid
	(1)	(2)	(3)	(4)
Senate	0.435	0.335	0.097	0.198
	(1.287)	(1.265)	(1.263)	(1.286)
Incumbent	7.180***	6.761***	6.100***	6.215***
	(1.439)	(1.414)	(1.412)	(1.437)
Experience	−4.179***	−4.229***	−4.611***	−4.587***
	(1.302)	(1.280)	(1.278)	(1.301)
Gender	−4.166***	−3.641***	−3.722***	−3.965***
	(0.907)	(0.891)	(0.890)	(0.906)
Republican	11.893***	11.503***	11.617***	11.578***
	(0.849)	(0.834)	(0.832)	(0.847)
Electoral Safety	−0.517	−0.680	−0.720	−0.687
	(0.458)	(0.450)	(0.449)	(0.457)
Total Tweets	2.429***	2.373***	2.375***	2.471***
	(0.419)	(0.412)	(0.411)	(0.419)
Constant	22.814***	21.397***	21.547***	22.314***
	(0.820)	(0.806)	(0.805)	(0.819)
Observations	1,573	1,573	1,573	1,573
R^2	0.163	0.156	0.159	0.156
Adjusted R^2	0.159	0.152	0.155	0.152
Note:				*p<0.1; **p<0.05; ***p<0.01

Table A.2

	Dependent variable: Average Message Ideology			
	Ideology Hand	Ideology Code1	Ideology Code2	Ideology Hybrid
	(1)	(2)	(3)	(4)
Senate	−0.002	−0.005	0.007	−0.004
	(0.016)	(0.016)	(0.016)	(0.016)
Incumbent	0.004	0.008	0.022	0.013
	(0.018)	(0.018)	(0.018)	(0.018)
Experience	−0.062***	−0.057***	−0.073***	−0.064***
	(0.016)	(0.016)	(0.016)	(0.016)
Gender	−0.071***	−0.065***	−0.090***	−0.088***
	(0.011)	(0.011)	(0.011)	(0.011)
Republican	0.448***	0.412***	0.423***	0.430***
	(0.010)	(0.011)	(0.011)	(0.011)
Electoral Safety	−0.011*	−0.010*	−0.014**	−0.017***
	(0.006)	(0.006)	(0.006)	(0.006)
Total Tweets	0.015***	0.017***	0.017***	0.016***
	(0.005)	(0.005)	(0.005)	(0.005)
Constant	0.813***	0.780***	0.853***	0.836***
	(0.010)	(0.010)	(0.010)	(0.010)
Observations	1,573	1,573	1,573	1,573
R^2	0.583	0.534	0.552	0.561
Adjusted R^2	0.581	0.532	0.550	0.559
Note:				*p<0.1; **p<0.05; ***p<0.01