ZeMKI Working Papers

ZeMKI Working Paper | No. 60
ISSN 2510-9855

Cornelius Puschmann, Hendrik Meyer

Gaging LLMs' strengths and weaknesses in political content analysis

Universität Bremen

ZeMKI Zentrum für Medien-, Kommunikations- und Informationsforschung

Universität Bremen | University of Bremen
ZeMKI, Zentrum für Medien-, Kommunikations- und Informationsforschung
Linzer Str. 4, 28359 Bremen, Germany, E-mail: zemki@uni-bremen.de
www.zemki.uni-bremen.de | www.zemki.uni-bremen.de/en/

Cornelius Puschmann (puschmann@uni-bremen.de)
Cornelius Puschmann is Professor of Communication and Media Studies with a focus on digital communication at ZeMKI and leader of the digital communication and information diversity (DCID) lab. From 2012 to 2016, he headed the DFG project “Networking, Visibility, Information? Usage motives of informal digital communication genres among scientists” at the Institute for Library and Information Science at Humboldt-Universität zu Berlin. Between 2014 and 2015, he held a professorship for communication science with a focus on digital communication at Zeppelin University Friedrichshafen. After conducting research at the Alexander von Humboldt Institute for Internet and Society as part of the project “Networks of Outrage: Mapping the emergence of new extremism in Europe” from March to September 2016, he moved to the Leibniz Institute for Media Research / Hans Bredow Institute at the University of Hamburg in 2016. Cornelius Puschmann has been a visiting scholar at the Oxford Internet Institute at the University of Oxford, the Berkman Klein Center for Internet and Society at Havard University and the Department of Media Studies at the University of Amsterdam, among others.

Hendrik Meyer (hendrik.meyer-1@uni-hamburg.de)
Hendrik Meyer has been a research associate at the Faculty of Economics and Social Sciences since May 2021. His research interests lie primarily in the study of structures and amplifications of media-networked political discourses - especially their supposed "polarisation". He is also interested in how the society negociates social media and platform regulations as well as public discourses on media technologies based on algorithms and so-called "artificial intelligence". Currently, he is also working on the data processing and analysis of the project "Climate Futures" in the Cluster of Excellence CLICCS and researching potential polarisation dynamics of (climate) political online discourses. In his dissertation, he analyses and classifies findings generated from this research through communication studies and network sociology.

Working Paper No. 60, June 2026

Published by ZeMKI, Centre for Media, Communication and Information Research, Linzer Str. 4, 28359 Bremen, Germany. The ZeMKI is a Central Research Unit of the University of Bremen.

ISSN: 2510-9855

Gaging LLMs' strengths and weaknesses in political content analysis

Abstract

Recent innovation and growth in the capabilities of generative artificial intelligence have opened options for the scholarship of political communication. While large language models (LLMs) have much to offer to improve content analysis workflows in political communication, there is as of yet no consensus regarding which options deliver the best results to the field and scale best computationally. This paper investigates the performance of five distinct classification models across three content analysis tasks of varying complexity: identifying political content, classifying political issues, and detecting stance towards climate policies and movements in news articles. Our automated coding is validated by 36,320 individual human coding decisions. We show that (political) task complexity is a crucial determinant for model selection. While simpler, efficient models perform well on lower to moderate complexity tasks, both proprietary and commercial LLMs exhibit superior performance on the nuanced stance detection task when appropriately instructed. These findings underscore the need to balance performance gains with computational costs and CO2 emissions, advocating for a "frugality approach" to automated coding.

Keywords

LLMs, content analysis, political communication, automated coding, model benchmarking, stance detection, task complexity, zero-shot classification, computational sustainability, methodological frugality

1 Introduction

Computational methods have played an important role in the study of political communication for at least a decade, depending on how broadly one defines the use of computational techniques (Theocharis & Jungherr, 2021; Windsor, 2021). Recent innovation and growth in the capabilities of generative artificial intelligence have opened options for the scholarship of political communication (Gil De Zúñiga et al., 2024), but also pose significant challenges to researchers, as they must weigh quality and efficiency of computational analysis with questions of transparency, technical complexity and cost (both financial and environmental). Automated content analysis in particular has made substantial progress in the course of only a few years, improving substantially in terms of capability, but also becoming much more complex (Chew et al., 2023; Gilardi et al., 2023; Kroon et al., 2024; Törnberg, 2024b). In particular, large encoder-decoder models (LLMs) relying on deep transfer learning and the transformer architecture facilitate computational analysis of complex and nuanced communicative concepts, such as media frames, incivility and stance, that were previously difficult to code reliably at a large scale (Laurer et al., 2024; Stoll, Wilms & Ziegele, 2023; Viehmann et al., 2023). Commercial LLMs such as GPT (Open AI) and Claude (Anthropic), as well as open source competitors such as Llama, Mistral and more recently DeepSeek, have demonstrated impressive capabilities as tools for media and communication research (Farjam et al., 2025; Gilardi et al., 2023; Törnberg, 2024b), though depending on the use case their accuracy has also been challenged (Bucher & Martini, 2024; Xu et al., 2024). The

capabilities of LLMs go beyond content analysis and include a diverse range of options, from the generation of stimulus material for experiments to serving as a partner in qualitative interviews (Zarouali et al., 2024). However there are also substantial challenges in utilizing large language models for analyzing content relevant to media and communication scholars. These can be subdivided into aspects of (1) validity, (2) reliability, (3) efficiency and – for proprietary models – (4) transparency. Developing a strategy that combines the strengths of LLMs with those of classic encoder-based transformers and matches computational resources with analytical needs thus seems desirable to advance the field methodologically (Kroon et al., 2024).

In this article, we follow the pioneering approach taken by Dun et al. (2021) who describe the combined use of dictionaries and supervised machine learning (SML) to study mass media coverage of policy changes in the U.S., but instead ask what different forms of automated coding based on LLMs have to offer for the study of political communication. Our approach is that of a straightforward methodological benchmark of different analytical strategies against one another, taking into account both the quality of coding and the resources needed to computationally administer them. Benchmarking different approaches in this fashion is also called for because the variety of available computational tools and models can lead to a paradox of choice, where researchers lack empirically grounded guidance on selecting the most appropriate model for their specific research question, dataset characteristics, and resource constraints. Substantially we argue that the complexity of the analytical task is key in selecting an appropriate classification model for automated content analysis. While sophisticated LLMs may offer superior performance on tasks requiring nuanced understanding and inferential reasoning, simpler, more efficient models can be equally, if not more, suitable for less complex tasks in studying political communication and political media. Performance must be critically weighed against computational and environmental costs associated with state-of-the-art artificial intelligence. Accordingly, this paper contributes to integrating a sustainability imperative as a critical dimension of model evaluation in computational communication research (van Atteveldt & Peng, 2018).

To investigate these issues, we compare the performance of *five distinct classification models*: three prominent LLMs (GPT-4o, Claude Sonnet 3.7, Llama 3.3 70b), as well as a zero-shot Natural Language Inference (NLI)-based model (Laurer et al., 2024), and a series of self-built few-shot fine-tuned sentence transformers with a classification head (Tunstall et al., 2022). These models are evaluated across three tasks common in communication research: the binary classification of news texts as political or non-political (*is political*), the binary classification of political issues within texts as present or absent (*issue*), and the multi-class detection of stance (supportive, opposed, or neutral) towards climate policy measures (*stance*). Rather than English-language data, which is widely used to train LLMs, we increase the relative difficulty by coding novel German-language data which in contrast to many popular English-language public datasets is certain to be unknown to the LLMs. Performance is assessed using a comprehensive set of metrics, and the findings are discussed in the context of task complexity and the inherent trade-offs between analytical performance, computational cost, and CO2 emissions. In the following, we first discuss recent developments in applying automated content analysis in political communication, as well as promises and limitations in the use of LLMs, and then proceed to model evaluation.

2 From bag-of-words to embeddings in political content analysis

Up to the late 2010s, the analysis of textual data in (computational) political communication research traditionally relied on bag-of-words (BoW) models, which represent texts as an unordered collection of words (Dun et al., 2021; Kroon et al., 2024). This approach, while computationally efficient and highly transparent, fundamentally disregards grammatical structure and the contextual meaning of words, which are critical for language comprehension. A partial remedy involves using short word sequences (n-grams) or phrases, but this technique fails to capture long-distance linguistic dependencies. Such limitations are particularly problematic when considering the degree to which digital communication from sources such as social media is variable, multilingual and departs from traditional stylistic norms. In the recent past, BoW has underpinned both dictionary-based methods and "naive" (i.e. non-pre-trained) supervised machine learning (Boumans & Trilling, 2016). Dictionary-based approaches utilize predefined word lists to classify texts (Soroka et al., 2015). However, their validity is highly context-dependent, as the meaning of words varies significantly across domains and usage situations (Puschmann et al., 2022). This specificity limits the generalizability of dictionaries, and their creation is susceptible to researcher bias. Consequently, applying pre-existing dictionaries to new or diverse contexts may yield poor results and can fail to capture latent or abstract theoretical constructs (González-Bailón & Paltoglou, 2015). Supervised machine learning (SML) using BoW representations offers an alternative that can better account for the specificities of a given dataset and reduce researcher bias (van Atteveldt et al., 2008). These models are trained on labeled data to learn which words map to specific categories. However, while SML often outperforms dictionaries, it also suffers from limited generalizability. A model trained on one corpus, such as newspaper articles, will perform poorly when applied to a different domain, like social media posts or political speeches (Kroon et al., 2024). Achieving high performance requires large, representative, and manually annotated training datasets, which are often prohibitively expensive and time-consuming to create for many researchers. Ultimately, both BoW-based approaches are vulnerable to variations in genre and style, potentially introducing systematic biases in analyses of diverse digital data. To overcome these limitations, the field has shifted towards text representations that capture deeper linguistic knowledge – word embeddings, which represent words as continuous numerical vectors rather than discrete, orthogonal categories. A further advance was realizing these embeddings could be learned from large, unlabeled text corpora, an early form of transfer learning. Subsequent developments, such as Convolutional and Recursive Neural Networks (CNNs and RNNs), enabled the creation of context-aware models that could learn complex patterns from word sequences, capturing both short- and long-distance dependencies and (among other advantages) facilitating cross-lingual analyses based on semantically similar constructs (Licht, 2023). This development towards (more) context-aware, transferable language knowledge has ultimately enabled the analysis of more complex and nuanced theoretical concepts relevant to political communication scholars.

The shift from a chiefly bag-of-words approach to one using word embeddings has been ongoing in (computational) media and communication research and political communication for several years and has led to concrete recent advances in substantive research. For example, Cochrane et al. (2022) evaluate whether speech transcripts can capture the emotional content of political speech by comparing human coding of video and text. To test automated methods, they compare several dictionary-based approaches and supervised machine learning models in their ability to predict human sentiment scores. The authors contribute the validation of a sentiment dictionary based on word embeddings (word2vec), which are generated from a large corpus of parliamentary debates and then used to calculate word sentiment based on proximity to positive and negative seed words. Similarly Müller et al. (2023) used semi-supervised machine learning

to analyze the explicit and implicit stigmatization of ethnic and religious groups in German news articles. Explicit stigmatization was measured using Latent Semantic Scaling (LSS), which expands a seed dictionary of "fear" and "admiration" words to score the sentiment of sentences containing group names. Implicit stigmatization is measured by calculating word embedding biases (using GloVe) to quantify a group's association with fear and admiration attribute words within the entire news corpus. Finally, Heseltine et al., (2025) combined a three-wave panel survey with web browser data from participants in the U.S. and Poland to measure exposure to partisan media and news about polarization. The authors identify exposure to partisan news by matching visited domains to lists of news outlets with pre-computed ideology scores. To measure exposure to news about polarization, they trained a BERT-based neural binary classifier on manually annotated news articles to automatically identify content covering political divides and conflict.

Examining these examples is noteworthy that in tandem with the shift from bag-of-words to word embeddings and from "naive" supervised machine learning to pretrained models, a simultaneous evolution of content analysis concepts subjected to automated classification has taken place, as evidenced in these studies: Concepts that are less theoretically salient but could be coded more reliably in the past (e.g. sentiment) have given way to those which are harder to code, but also more relevant conceptually (racial stereotypes, perceptions of polarization). This shift resonates with the ability of LLMs to capture more complex and nuanced constructs.

3 Promises of coding with LLMs

LLMs such as GPT (Open AI), Gemini (Google) and Claude (Anthropic), as well as a growing list of open source competitors, such as Llama, Mistral and more recently DeepSeek, have the unique ability to take context into account for inference and can be instructed via extensive prompting (Farjam et al., 2025). Based on elaborate prompts, these models are able to zero-shot classify new text examples with impressive quality (Brown et al., 2020) .

Recent comparative analyses have demonstrated their performance and adaptability in a variety of Natural Language Processing tasks, such as sentiment analysis, offensive language detection, stance detection, and accusation detection (Corral et al., 2024; Fields et al., 2024). The use of LLMs reduces the need for large labeled datasets, which are costly and time-consuming to produce (Kuzman et al., 2023). This efficiency makes them attractive to researchers and practitioners who wish to scale up their text analysis efforts. LLMs have demonstrated remarkable capabilities in zero-shot and few-shot learning scenarios, where they can perform classification tasks without explicit training on the specific dataset (Brown et al., 2020). For example Törnberg (2024a) found ChatGPT to outperform human annotators of social media messages in terms of speed and, in some cases, accuracy and reliability of annotation tasks. In the same vein, (Gilardi et al., 2023) found ChatGPT's zero-shot accuracy exceeding that of crowd workers (via Amazon Mechanical Turk) by an average of about 25% across four different datasets and various annotation tasks, suggesting that LLMs can produce more accurate annotations than crowd workers (but possibly not trained annotators) for certain problems. The authors extended this good level of performance to relatively difficult tasks, but noted that detailed prompts and the use of examples were of key importance for achieving good results. It has also been shown that LLMs are capable of identifying complex concepts, such as sexist or misogynistic language and claims of hypocrisy, as well as classifying text into categories such as hate speech, misinformation, and news credibility (Corral et al., 2024; Hoes et al., 2023; Huang et al., 2023; Plaza-del-arco et al., 2023; Yang & Menczer, 2025).

Accordingly, LLM-based classification approaches have shown promise in automating labour-intensive tasks such as text annotation and classification, which traditionally

require significant human effort (Kroon et al., 2024). As demonstrated by Farjam et al. (2025), LLMs not only perform many classification tasks reliably but also reveal underlying ambiguities in coding schemes and conceptual definitions. When prompted to identify latent or complex concepts with vague or inconsistently operationalized categories, LLMs tend to produce unstable or conflicting outputs. Rather than indicating model failure, such discrepancies may signal unresolved theoretical tensions within the coding framework itself. In this way, LLMs can function as diagnostic tools that compel researchers to confront and refine the conceptual boundaries of their categories. The result is a tool that not only enables large-scale coding but also pressures researchers to define their analytical categories more explicitly and consistently. Another, somewhat less obvious advantage of using LLMs (or more generally zero-shot classification) may be that both unintentional overfitting on even intentional strategies that inflate performance measures in classical supervised machine learning can be largely ruled out. Performance on out of sample data of SML (supervised machine learning) classifiers is often weaker than within sample test results suggest and researchers who seek to deliberately game the performance indicators of their classifiers have ample options to do this, as Stoll (2023) demonstrates in a thought-provoking methodological contribution on how this gaming is done in practice. Therefore zero-shot approaches with transparent prompting may not only deliver superficially good-looking but also more robust results.

4 Challenges of coding with LLMs

In spite of their clear potential, there are also substantial challenges in utilizing large language models for analyzing content relevant to media and communication scholars, namely their overall (1) *validity*, (2) *reliability*, (3) *efficiency* and – for proprietary models – (4) *transparency*. In the following we briefly discuss these issues, once by one, to then propose an alternative

(1) Given their relative recency, it is unsurprising that results on the base validity of LLMs for standardized content analysis tasks such as identifying issues, media frames, sentiment and stance vary widely. Ziems et al. (2024) found that on classification tasks across multiple computational social science benchmarks LLMs achieved "fair levels of agreement with humans." However, their performance was often not high enough to completely replace human annotation. Similarly, Rogers and Zhang (2024) observed that LLMs used for analyzing social media content related to the Russia-Ukraine war exhibited a "bias toward neutrality" when compared to human coders. This bias was evident in the tonality of the labeling and the tendency of LLMs to label texts as neutral more frequently than manual classification. Tekumalla and Banda (2023) assessed LLMs for annotating COVID-19 vaccine-related tweets and found that GPT models performed well in zero-shot labeling, but that at larger scales, misclassifications could become problematic if the tasks that incorporate are critical (e.g. in healthcare contexts). Ziems et al. (2024) also observed that for structured parsing tasks like event extraction, code-instructed models like OpenAI's models performed well, but less so for classification. (Törnberg, 2024a) in this context emphasizes that the ostensible simplicity of LLMs can be misleading, as they can be prone to bias, misunderstandings, and unreliable results. He also argues for a structured, directed, and formalized approach to using LLMs for text annotation, including rigorous (human) model validation, as undesirable behaviors can manifest suddenly and without explanation, leading to the second problem, lack of consistency.

(2) In addition to validity (in the sense of *quality of coding*), a second, separate problem concerns the robustness of classification results when using LLMs for coding. A comparison of model outputs with human-annotated reference data

remains essential and may fail at closer examination. For example Reiss (2023) raised concerns about their reliability, consistency, and reproducibility — especially when these models are applied to zero-shot classification tasks without supervised training. Despite their capabilities, LLMs are — by design and architecture — non-deterministic, meaning that identical inputs may yield varying outputs due to the randomness embedded in the generation process. This non-deterministic behavior of LLMs, also mentioned by Tekumalla and Banda (2023) necessitates caution on the part of the researchers to ensure consistent and reliable results. Reiss (2023) highlighted this issue by showing that repetitions of tasks using prompts/model parameters with small or no changes can lead to inconsistent classification results. In his study of classifying website text into news and non-news categories, the author found that the consistency of ChatGPT's results often fell below standard thresholds for reliability. Factors such as temperature settings — a parameter that controls the randomness of the model's output — significantly affected consistency. Higher temperatures led to more variable outputs, reducing reliability. However, even when repeating the same input with the same parameters across multiple iterations, the results varied, undermining the reliability of the classification. This inconsistency presents a challenge for tasks that require a high degree of reliability and reproducibility which was also found by other researchers: A recent study by Spirling (2023) found results in standardized CA to vary considerably from one run to another due to dynamicity hard-built into generative models, calling into question their ability to produce reliable and reproducible results, a key criterium in content analysis. Additionally, the way prompts are structured can have a significant impact on model performance. Both Reiss (2023) with regards to content analysis and Bisbee et al. (2024) in the context of survey research found that minor prompting differences that made no noticeable differences to humans lead to differences in results. Slight changes in the wording of prompts could lead to different classifications (Argyle et al., 2023; Furniturewala et al., 2024; Reiss, 2023). Thus, “prompt brittleness” and subtle or underlying language seem to be a limitation of such classification models (Ceron et al., 2024). Such issues may stem from political bias in the training data, hallucination, and other flaws introduced through varying architecture of different LLMs and the data it used, as "context matters for [identifying] a complex semantic concept" (Corral et al., 2024). Gilardi et al. (2023) also emphasize the need for further research to understand the capabilities of LLMs in a broader range of contexts, including the implementation of different prompting strategies and conclude that validity of LLM-based content analysis is contingent on careful task design and prompting.

(3) In addition to these challenges, LLMs are computationally expensive to deploy at scale, have comparatively long inference times per analyzed unit (in particular so called “reasoning” models) and require a larger expenditure of energy and consequently also cause greater CO2 emissions than simpler models. Samsi et al. (2023) evaluated the environmental impact of large language models by benchmarking the energy costs associated with inference, relying on Meta AI's LLaMA models. They found notable differences between processor architectures, highlighting that for large models requiring more than two GPUs, energy consumption even for inference can be substantial. Luccioni et al. (2023) adopted a more comprehensive life cycle assessment approach to evaluate the carbon footprint of the BLOOM model. While their study encompassed the carbon emissions from training, including embodied, dynamic, and idle power consumption, they also dedicated a portion to examining the energy requirements and carbon emissions of BLOOM inference via a real-time API on Google Cloud Platform. They observed that a substantial fraction of the total energy consumed

(around 75%) was attributed to GPU usage, even when the model was idle, highlighting the energy cost of simply keeping a large model loaded in memory for potential inference requests. Based on the energy consumption and the carbon intensity of the Google Cloud region, Luccioni et al. (2023) estimated the carbon emissions of this specific BLOOM API deployment to be approximately 19 kg of CO2 per day, equivalent to driving roughly 80 kilometers with a combustion-powered car. Rillig et al. (2023) offered a more conceptual evaluation of the environmental impact of large language models, identifying increased energy consumption for both the training and inference of LLMs as a direct negative consequence. The authors also highlighted that the overall carbon footprint of large language models is contingent on both their size and associated computational cost and degree of their adoption. For example, inference fully replacing online search could have significant environmental consequences (Jegham et al., 2025).

Finally, a series of arguments have been made against the use of proprietary models due to their lack of transparency. Since the inner workings of proprietary LLMs are unknown to users and may rapidly change over time, it has been argued that they should not be used in social science unless they clearly represent the best alternative in terms of performance when compared with open source alternatives. Bisbee et al. (2024) found that the distribution of responses to the same prompt changed significantly over a 3-month period due to alterations in the underlying algorithm, hampering reproducibility. Palmer et al. (2024) also note that companies can substantially alter or even abandon their algorithms at any time, also affecting the reproduction of earlier findings. The difficulty in understanding model behavior and biases is another key problem. The authors emphasize that being able to audit the data used by a system helps researchers understand its behavior – an issue also highlighted by Baack (2024) in his critical assessment of common crawl. Bisbee et al. (2024) discuss how elements of ChatGPT are biased against certain types of responses due to reinforcement learning with human feedback (RLHF), but note that the specifics of these biases are not fully transparent. Ethical concerns are also raised by the reliance on predictions from models with unknown methods and the authors argue that using such models as a substitute for asking humans their opinions goes against the origins of public opinion research. In this vein, Palmer et al. (2024) advocate for scientists to explicitly justify their use of proprietary models in research. Making the choice and timing explicit could help to informally "version" closed models by documenting their capabilities at a given time and encourage the adoption of more open technologies, a view also promoted emphatically by Spirling and colleagues (2023).

5 Benchmarking different models across tasks in political communication

In the following, we describe the results of a benchmarking of five different models against three different content analysis tasks widely applicable to political communication. Below we first describe the human labeling of the three datasets (n = 36,320 coding decisions)[1].

Tasks

is_political: A sample of 1,000 German-language online news texts was manually coded as either "political" or "non-political" by five human coders based on detailed coding instructions. A good level of intercoder reliability was achieved (Krippendorff's α = 0.83).

[1] All validation data and R code used in this paper are available anonymously via OSF: https://osf.io/ac4pv/?view_only=46269dea1b8d47bea9cbb55ce2fd3bae

The consensus cases with agreeing labels from all coders (n = 944) were subsequently used for validation (and, in the case of the SetFit classifier, for training, in a conventional 80/20 training test split). This task represents a relatively low-complexity binary classification, often relying on surface-level lexical and thematic cues, such as the mentioning of politicians, parties and political institutions.

issue_classification: A sample of 5,000 German-language online news articles was manually coded as being associated with one of five political issues (climate change, gendered language, inflation, transgender rights and the invasion of Ukraine) by two coders each based on detailed coding instructions. Coding was binary for each issue separately (rather than nominal) based on a balanced sample that had been previously stratified using BERTopic. A good level of human agreement was achieved across issues (climate α = 0.8, gender α = 0.95, inflation α = 0.99, transgender, α = 0.85, ukraine α = 0.96). Again, the consensus cases were used for validation (and, in the case of the SetFit classifier, for training). This yielded stratified samples of slightly varying size for each issue (climate: 452, gendered language: 492, inflation: 494, transgender rights: 463, ukraine: 489). This too is a binary classification task, in this case of moderate complexity, requiring a deeper thematic understanding of the text in order to infer thematic associations.

stance_detection: A sample of 264 German-language online news paragraphs discussing climate policy measures (introduction of a national speed limit and mandating the use of heat pumps in private households) and climate protest movements (Fridays for Future and The Last Generation) was manually coded by five human coders based on detailed coding instructions for the stance expressed towards those measures and movements in the article, specifically as "supportive," "opposed," or "neutral" (protest: 89, speed limit: 92, heat pumps: 83). This multi-class classification task represents a higher level of complexity, demanding nuanced interpretation of opinions, attitudes, and the relationship between the expressed view and the specific policy target or protest movement. As a result of the relative difficulty of the task, the small sample size and the relative conservativeness of Krippendorff's alpha in penalizing disagreement, only mediocre results were achieved in terms of reliability (protest α = .57, speed limit α = .62, heat pumps α = .54). However this problem was partly alleviated by comparing the results of automated coding with a majority vote of human coders where applicable, which was treated as the ground truth.

Data

The textual data for these tasks were sourced from two large corpora of German-language online news articles. The first dataset was collected in the context of a large-scale panel study in which participant browser activity was passively tracked. This tracking data was collected in Germany between March 2022 and December 2023. An independent market research organization, Bilendi SA recruited and supplied the panel. The panel provided recorded browsing activity on both mobile and desktop devices, logging the visited complete URL, time of visit, duration of visit and the device being utilized. Tracking was active for up to 24 hours a day, contingent on the participants' choice to disable it for periods of time if they so choose. This resulted in a dataset of 1.1 mio. articles from a variety of different online news domains (Puschmann et al., 2023).

The second dataset is based on a large-scale content analysis of German news articles focusing on specific climate change-related debates. These include discussions on climate protests, ecological transformation involving heating infrastructure, and measures related to traffic infrastructure. The sample comprises 17 news outlets selected with an emphasis on ideological diversity, encompassing a range of source types—including daily newspapers

and magazines—available in both print and online formats. A comprehensive approach was taken to determine the most relevant outlets. The selection was based on a composite ranking that considered relevance across print, online, and social media or blog platforms (drawing on data from *Reuters Digital News Reports* and *Arbeitsgemeinschaft Medien-Analyse*). Articles from outlets with both print and online versions were merged.

The data was retrieved from the Factiva database using targeted search queries related to climate protests, heating measures, and traffic infrastructure being published between January 2022 and December 2023. The resulting corpus comprised more than 36,000 articles.

Models

We compare five models, representing different architectural paradigms and training approaches. The inclusion of both state-of-the-art zero-shot and highly efficient few-shot specialized models provides critical baselines against the LLMs, moving beyond a simplistic comparison to older methods and offering a more nuanced evaluation against contemporary, efficient alternatives.

Large Language Models (LLMs):

GPT-4o (OpenAI)
Claude Sonnet 3.7 (Anthropic)
Llama 3.3 70b (Meta)

Internal consistency (i.e. reliability of classification across runs) of LLMs was tested by repeating the *is_political* task a total of three times with GPT and Claude. In 98% (GPT-4o) and 99% (Claude Sonnet) of all cases, the model classified each item identically across all three runs, suggesting a good level of internal reliability. We found this to be reassuring in light of the fact that humans too are likely to alternate in their judgment across runs (Bojic et al., 2025).

Zero-Shot NLI-based Model:

This model (ml_zeroshot) was based on Moritz Laurer's bge-m3-zeroshot-v2.0.44 (Laurer et al., 2024). It utilizes the BGE-M3 sentence embedding model within a Natural Language Inference (NLI) framework to perform zero-shot classification, i.e. classifies texts into predefined categories without having been explicitly trained on examples from those specific categories for the given task.

Few-Shot Fine-tuned Sentence Transformer:

SetFit refers to a series of models we fitted where a sentence transformer (from the Sentence-BERT family, in this case paraphrase-multilingual-MiniLM-L12-v2) was fine-tuned for each of the three tasks using the SetFit framework. SetFit is designed for high performance with very few labeled examples per class, making it an efficient approach for creating task-specific classifiers (Tunstall et al., 2022).

Performance Metrics

A comprehensive suite of standard classification metrics was employed to evaluate model performance robustly:

Accuracy: The proportion of correctly classified instances.

Precision: The proportion of true positive predictions among all positive predictions made by the model (TP / (TP + FP)). For the multi-class task, this is macro-averaged.

Recall: The proportion of actual positives that were correctly identified by the model (TP / (TP + FN)), also known as sensitivity. For the multi-class task, this is macro-averaged.

F1-score (f_meas): The harmonic mean of precision and recall. For the multi-class task (stance_detection), this is reported as macro-averaged F1-score, which calculates the F1 for each class independently and then averages them, treating all classes equally. For the binary tasks is_political and issue, it is the standard F1-score.

Cohen's Kappa (kap): A statistic that measures inter-rater agreement for categorical items, adjusted for agreement occurring by chance. It is used here to assess the agreement between model predictions and ground truth labels. κ can be calculated directly from a confusion matrix:

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$$

Matthews Correlation Coefficient (mcc): A balanced measure of classification quality, particularly useful even if the classes are of very different sizes. It returns a value between -1 and +1. The MCC can be calculated directly from a confusion matrix:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

We employ such a range of measurements because (for example) F1 has been criticized as an imperfect measure that produces overly optimistic results in cases of class imbalance, while other (better) measures, such as MCC are not universally used.

Classification results

The results of the comparative evaluation of the five classification models across the three automated content analysis tasks are summarized visually in Figure 1 and detailed comprehensively in Table 1.

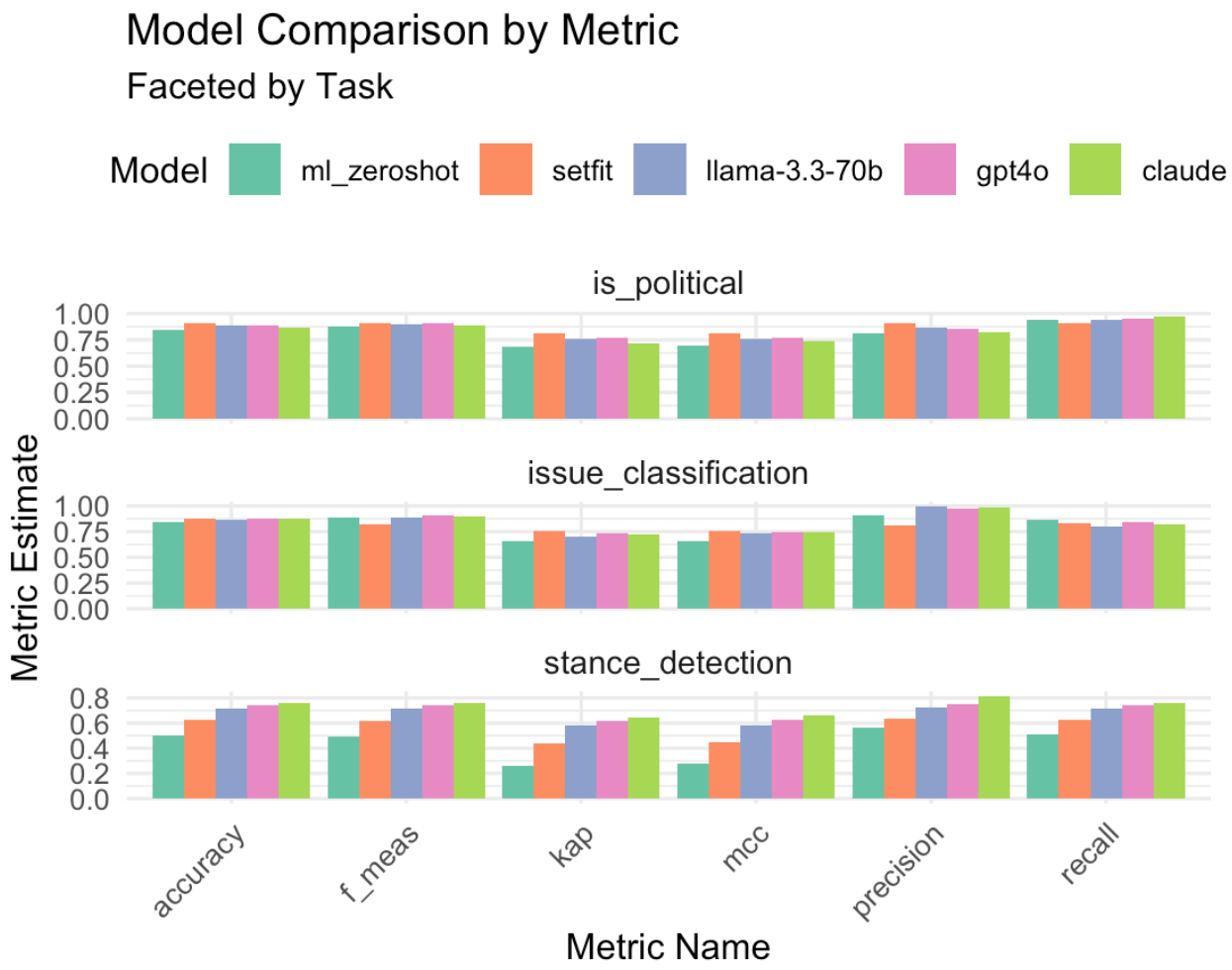

Figure 1: Model performance comparison by metric, faceted by classification task.

The figure provides a high-level overview of model performance. Across tasks, different patterns emerge. For *is political* and *issue* classification, the performance for SetFit and the LLMs appear comparable, while ml_zeroshot also shows competitive results. However, for *stance*, the LLMs (light green, pink, and dark blue bars) generally exhibit substantially higher scores across metrics compared to ml_zeroshot (mint green) and SetFit (orange).

Task	Model	Accuracy	F1-score (binary/ macro)	Kappa	MCC	Precision (binary/ macro)	Recall (binary/ macro)
is_political	gpt4o	0.89	0.91	0.77	0.77	0.86	0.96
	claude	0.87	0.89	0.72	0.74	0.82	0.97
	llama-3.3-70b	0.88	0.9	0.76	0.76	0.87	0.94
	SetFit	0.91	0.91	0.81	0.81	0.91	0.91
	ml_zeroshot	0.85	0.88	0.68	0.7	0.82	0.95
issue	gpt4o	0.88	0.9	0.73	0.75	0.98	0.84
	claude	0.87	0.9	0.73	0.75	0.99	0.82
	llama-3.3-70b	0.86	0.89	0.71	0.74	0.99	0.8
	SetFit	0.88	0.82	0.75	0.75	0.81	0.83
	ml_zeroshot	0.85	0.89	0.65	0.65	0.91	0.87
stance	gpt4o	0.75	0.74	0.62	0.62	0.75	0.75
	claude	0.76	0.76	0.64	0.67	0.81	0.76
	llama-3.3-70b	0.72	0.72	0.58	0.58	0.73	0.72
	SetFit	0.62	0.62	0.43	0.44	0.63	0.62
	ml_zeroshot	0.5	0.5	0.26	0.27	0.56	0.51

Table 1: Comprehensive performance metrics for models across tasks.

Task 1: is_political (Low Complexity)

On this binary classification task, all models performed well. The SetFit model achieved the highest accuracy (0.91) and F1-score (0.91), closely followed by the LLMs (GPT4o: Acc 0.89, F1 0.91; Llama-3.3-70b: Acc 0.88, F1 0.90). ml_zeroshot also demonstrated good competence (Acc 0.85, F1 0.88). Claude showed high recall (0.97), indicating it was effective at identifying most political texts, though its precision was slightly lower than other LLMs. The performance differences among the top models were marginal, suggesting that for this relatively straightforward task, the advanced capabilities of LLMs did not confer a decisive advantage over the efficiently fine-tuned SetFit or even the zero-shot ml_zeroshot.

Task 2: issue_classification (Moderate Complexity)

For this binary classification task, SetFit again demonstrated robust performance (Acc 0.88, F1_macro 0.82). The LLMs also performed strongly, with GPT4o (Acc 0.88, F1_binary 0.90) and Claude (Acc 0.87, F1_binary 0.90) leading slightly. Llama 3.3 70b showed very high precision (0.99) but lower recall (0.80), suggesting it was very accurate when it did classify an issue but missed more instances compared to others. The ml_zeroshot model was also competitive (Acc 0.85, F1_binary 0.89). The results indicate that while LLMs perform well, a well-tuned SetFit model remains a highly viable option.

Task 3: stance_detection (High Complexity)

This task showed the most significant performance differentiation among the models. The LLMs clearly outperformed SetFit and ml_zeroshot across all metrics. Claude Sonnet 3.7 achieved the highest F1-macro (0.76), MCC (0.67), and precision_macro (0.81). GPT-4o followed closely (F1_macro 0.74, Acc 0.75). Llama 3.3 70b also performed competently (F1_macro 0.72). In contrast, SetFit achieved an F1-macro of 0.62, and ml_zeroshot scored considerably lower with an F1-macro of 0.50. This substantial gap highlights the LLMs' superior ability to handle the nuanced understanding of opinion and context required for accurate stance detection.

The empirical results confirm our assumptions: the complexity of the analytical task is a key factor in determining relative model performance. For the less complex is_political and moderately complex issue_classification tasks, the specialized models (SetFit and ml_zeroshot) demonstrated performance levels that were often comparable, and in some cases (like SetFit for is_political), superior to the general-purpose LLMs. The performance of these efficient models might be considered "good enough" or even excellent for many research applications, especially when their lower computational and environmental costs are factored in. For instance, an F1-score around 0.90 for is_political by SetFit and ml_zeroshot is likely sufficient for robust filtering. Similarly, an F1-macro of ~0.82 by SetFit for issue_classification indicates strong thematic categorization. For the highly complex stance_detection task, the LLMs (GPT-4o, Claude Sonnet 3.7, Llama 3.3 70b) exhibited a clear and substantial performance advantage over both SetFit and ml_zeroshot. The F1-macro scores for LLMs ranged from approximately 0.72 to 0.76, whereas SetFit achieved ~0.62 and ml_zeroshot only achieved ~0.50. For a task like stance detection, which often underpins sensitive analyses of public opinion or political division, the performance of the smaller models might fall below an acceptable threshold for drawing reliable scholarly conclusions, making the superior (if costly) performance of LLMs more compelling. Even within model tiers, specific metric strengths emerge. For instance, in stance detection, Claude Sonnet 3.7's particularly high precision_macro (0.81) and MCC (0.67) suggest it excels at making confident positive classifications with fewer false positives, a valuable trait when the cost of such errors is high. This more fine-grained view allows us to highlight task-specific operational advantages that should be relevant to media and communication researchers interested in a range of different constructs in content analysis.

Cost and energy consumption comparison

The estimates described below are derived from the models utilized in the analysis combined with published pricing and energy benchmarks.

Task	Complexity	Sample (consensus cases)
is_political	Low	944 items
issue_classification	Moderate	2,390 items (across 5 issues: climate 452, gendered language 492, inflation 494, transgender rights 463, Ukraine 489)
stance_detection	High	264 items

Table 2: Classification task sample sizes

Reliability testing: GPT-4o and Claude Sonnet 3.7 each ran the *is_political* task three times to assess internal consistency. This triples the API call count for that task for those two models.

Token estimation per API call: The German news texts average approximately 300–600 words. At a conservative estimate of ~1.3 tokens per word for German (compound words inflate token counts relative to English), a 400-word article yields roughly 520 content tokens. Adding the system prompt and task instructions (estimated at 300–400 tokens for the detailed zero-shot prompts described in the paper), and a minimal output of ~10–15 tokens (a classification label), the per-call estimates are:

Task	Input tokens (text + prompt)	Output tokens
is_political	~820	~10
issue_classification	~870	~10
stance_detection	~550	~15

Table 3: Input/output token calculation per task

Table 4 below summarises total API call counts, token volumes, and estimated monetary costs. Prices reflect the API rate structure in place in early 2025, when the study was conducted: GPT-4o at $2.50/1M input and $10.00/1M output tokens; Claude Sonnet 3.7 at $3.00/1M input and $15.00/1M output; and Llama 3.3 70b via a commercial hosting provider (e.g., Groq or Together.ai) at approximately $0.59/1M input and $0.79/1M output.

Model	API calls	Input tokens	Output tokens	Estimated cost (USD)
GPT-4o	5,486	4,546,740	56,180	~$11.93
Claude Sonnet 3.7	5,486	4,546,740	56,180	~$14.48
Llama 3.3 70b	3,598	2,998,580	37,300	~$1.80
Total (LLMs)	**14,570**	**12,092,060**	**149,660**	**~$28.21**

Table 4: Model pricing estimates

The aggregate cost for the validation sample is modest at approximately $28 USD. However, the scaling implications are substantial: applying GPT-4o to the full corpus of 1.1 million articles would, under the same token assumptions, cost on the order of $2,300–$2,500 USD for that model alone. The locally run models (SetFit, ml_zeroshot) incur no API costs and scale without marginal expense.

Energy estimates follow the methodology of Epoch AI (2025) and Jegham et al. (2025), which distinguish between the input (prefill) and output (decode) phases of inference. The prefill phase is approximately ten times more energy-efficient per token than the decode phase. For GPT-4o-class models (estimated ~100B active parameters on H100 hardware), this yields approximately 0.00006 Wh per input token and 0.0006 Wh per output token. For Llama 3.3 70b, a dense 70B-parameter model, these figures are approximately halved. A cloud data centre carbon intensity of 0.2 kg CO_2/kWh is assumed.

Model	Energy (Wh)	CO_2 equivalent (g)
GPT-4o	~306.5 Wh	~61.3 g
Claude Sonnet 3.7	~306.5 Wh	~61.3 g
Llama 3.3 70b	~101.2 Wh	~20.2 g
Total (LLMs)	**~714 Wh**	**~143 g**

Table 5: Estimated inference energy consumption

For reference, the locally run models in the study consume a small fraction of this. Based on published benchmarks for task-specific fine-tuned models, the SetFit classifier (paraphrase-multilingual-MiniLM-L12-v2, ~33M parameters) consumes an estimated ~0.01 Wh for the full 3,598-item validation set, and the ml_zeroshot NLI model (bge-m3-zeroshot-v2.0.44, ~570M parameters) approximately ~0.11 Wh. The energy differential

between a large proprietary LLM and a fine-tuned sentence transformer for the same classification workload is therefore on the order of 3 to 4 orders of magnitude or roughly 2,800× to 28,000× more energy for GPT-4o than for SetFit.

5 Discussion

Our aim with this paper has been to benchmark different approaches to automated content analysis in political communication utilizing large language models. To this end we have compared five distinct classification models—GPT-4o, Claude Sonnet 3.7, Llama 3.3 70b, ml_zeroshot, and SetFit—across three automated content analysis tasks common in communication and media research. Our findings underscore that task complexity is a paramount factor in model selection. We find that large language models are well-suited to coding for complex constructs such as stance, where they are able to deliver satisfactory results even in the face of categories that are challenging to reliably code for human experts. However, these advantages must be weighed with their relatively high computation cost, and the lack of transparency and reproducibility that applies in particular to proprietary models. We emphasize that some relatively simple conceptual categories that are nonetheless relevant to political communication research can be coded reliably and with great computational efficiency using older and simpler models that run on conventional hardware, obviating the need for expensive kit or proprietary AI APIs.

In other words, our results demonstrate that for tasks of low to moderate complexity, such as identifying political content or categorizing texts into general issue areas, computationally leaner and more efficient models like the few-shot fine-tuned SetFit and the zero-shot NLI-based ml_zeroshot achieve performance levels that are not only adequate but comparable to those of large language models. In contrast, for the more nuanced and inferentially demanding task of stance_detection, the LLMs showcased a clear and substantial performance advantage, highlighting their advanced capabilities in interpreting subtle linguistic cues and complex contexts, especially when provided with in-depth explanations and concrete examples.

This said, performance metrics alone paint an incomplete picture. The decision-making process for model selection in computational social research must also integrate the crucial dimensions of computational cost and environmental sustainability. LLMs, while powerful, come with significantly higher resource demands and a larger carbon footprint. This research, therefore, champions a principle of "methodological frugality"—an approach that advocates for using the simplest, most efficient model that can adequately and robustly address the research question at hand, rather than defaulting to the largest or most computationally intensive model available. This aligns with the scientific principle of parsimony and promotes more sustainable and responsible research practices within computational communication science.

Our study also comes with several limitations. Coding for stances in relation to policy issues in mainstream newswriting is difficult because professional journalists generally refrain from taking clear sides or expressing very strong personal stances towards issues or political actors. By contrast, stance in newswriting is generally subtle and subject to interpretation. Considering these conceptual challenges, the performance of LLMs is promising, but we should also not assume that LLMs can resolve them for us. Additionally, we achieve only moderate human intercoder-agreement in the stance task, which has downstream consequences for the quality of the LLM-based classification. As always, high-quality human labeling is the prerequisite for successful automation. At the same time, our classification of issues as political or non-political as well as our definitions of particular politically salient issues, while based on a detailed codebook, can be challenged on the grounds of being relatively conservative. For example, content was classified as

political when explicitly mentioning political actors or topics, usually with a bias towards the political institutional political system, rather than a wide array of cultural or personal questions that can be considered deeply political.

The paradox of choice presented by the expanding array of AI tools can be navigated by a careful consideration of the specific analytical task, a realistic assessment of required performance thresholds, and an ethical awareness of the resource implications. As automated content analysis continues to transform inquiry in communication and social sciences, the ultimate goal remains the enhancement of social scientific understanding through methods that are not only powerful and innovative but also robust, efficient, and ethically sound. Future research should continue to build upon these findings, developing clearer guidelines and fostering a culture of informed and responsible AI adoption within these vital fields of study.

Our final warning echoes that of other colleagues who urge researchers not to believe too blindly in the promises particularly of proprietary LLM providers for content analysis and to be critical when it comes to coding and labeling through LLMs. The dictum of Grimmer and Stewart (2013) to "validate, validate, validate" still rings true even well over a decade after it was originally formulated.

6 References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Baack, S. (2024). A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2199–2208. https://doi.org/10.1145/3630106.3659033

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, *32*(4), 401–416. https://doi.org/10.1017/pan.2024.5

Bojic, L., Zagovora, O., Zelenkauskaite, A., Vukovic, V., Cabarkapa, M., Jerkovic, S. V., & Jovančevic, A. (2025). *Evaluating Large Language Models Against Human Annotators in Latent Content Analysis: Sentiment, Political Leaning, Emotional Intensity, and Sarcasm* (No. arXiv:2501.02532). arXiv. https://doi.org/10.48550/arXiv.2501.02532

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit. *Digital Journalism*, *4*(1), 8–23. https://doi.org/10.1080/21670811.2015.1096598

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Bucher, M. J. J., & Martini, M. (2024). *Fine-Tuned "Small" LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification* (No. arXiv:2406.08660). arXiv. https://doi.org/10.48550/arXiv.2406.08660

Ceron, T., Falk, N., Barić, A., Nikolaev, D., & Padó, S. (2024). Beyond Prompt Brittleness: Evaluating the Reliability and Consistency of Political Worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, *12*, 1378–1400. https://doi.org/10.1162/tacl_a_00710

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). *LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding* (No. arXiv:2306.14924). arXiv. https://doi.org/10.48550/arXiv.2306.14924

Cochrane, C., Rheault, L., Godbout, J.-F., Whyte, T., Wong, M. W.-C., & Borwein, S. (2022). The Automatic Analysis of Emotion in Political Speech Based on Transcripts. *Political Communication*, *39*(1), 98–121. https://doi.org/10.1080/10584609.2021.1952497

Corral, P. G., Green, A., Meyer, H., Stoll, A., Yan, X., & Reuver, M. (2024). *A Few Hypocrites: Few-Shot Learning and Subtype Definitions for Detecting Hypocrisy Accusations in Online Climate Change Debates* (No. arXiv:2409.16807). arXiv. https://doi.org/10.48550/arXiv.2409.16807

Dun, L., Soroka, S., & Wlezien, C. (2021). Dictionaries, Supervised Learning, and Media Coverage of Public Policy. *Political Communication*, *38*(1–2), 140–158. https://doi.org/10.1080/10584609.2020.1763529

Epoch AI. (2025, February 7). How much energy does ChatGPT use? https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use

Farjam, M., Meyer, H., & Lohkamp, M. (2025). A Practical Guide and Case Study on How to Instruct LLMs for Automated Coding During Content Analysis. *Social Science Computer Review*, 08944393251349541. https://doi.org/10.1177/08944393251349541

Fields, J., Chovanec, K., & Madiraju, P. (2024). A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe? *IEEE Access*, *12*, 6518–6531. https://doi.org/10.1109/ACCESS.2024.3349952

Furniturewala, S., Jandial, S., Java, A., Banerjee, P., Shahid, S., Bhatia, S., & Jaidka, K. (2024). *Thinking Fair and Slow: On the Efficacy of Structured Prompts for Debiasing Language Models* (No. arXiv:2405.10431). arXiv. https://doi.org/10.48550/arXiv.2405.10431

Gil De Zúñiga, H., Goyanes, M., & Durotoye, T. (2024). A Scholarly Definition of Artificial Intelligence (AI): Advancing AI as a Conceptual Framework in Communication Research. *Political Communication*, *41*(2), 317–334. https://doi.org/10.1080/10584609.2023.2290497

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

González-Bailón, S., & Paltoglou, G. (2015). Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 95–107. https://doi.org/10.1177/0002716215569192

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Heseltine, M., von Hohenberg, B. C., Menchen-Trevino, E., Gackowski, T., & Wojcieszak, M. (2025). Effects of Over-Time Exposure to Partisan Media and Coverage of Polarization on Perceived Polarization. *Political Communication*, *42*(3), 432–453. https://doi.org/10.1080/10584609.2024.2423093

Hoes, E., Altay, S., & Bermeo, J. (2023). *Leveraging ChatGPT for Efficient Fact-Checking*. https://doi.org/10.31234/osf.io/qnjkf

Huang, F., Kwak, H., & An, J. (2023). *Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech* (No. arXiv:2302.07736). arXiv. https://doi.org/10.48550/arXiv.2302.07736

Jegham, N., Abdelatti, M., Elmoubarki, L., & Hendawi, A. (2025). *How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference* (No. arXiv:2505.09598). arXiv. https://doi.org/10.48550/arXiv.2505.09598

Kroon, A., Welbers ,Kasper, Trilling ,Damian, & and van Atteveldt, W. (2024). Advancing Automated Content Analysis for a New Era of Media Effects Research: The Key Role of Transfer Learning. *Communication Methods and Measures*, *18*(2), 142–162. https://doi.org/10.1080/19312458.2023.2261372

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). *ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification* (No. arXiv:2303.03953). arXiv. https://doi.org/10.48550/arXiv.2303.03953

Laurer, M., Atteveldt, W. van, Casas, A., & Welbers, K. (2024). Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, *32*(1), 84–100. https://doi.org/10.1017/pan.2023.20

Licht, H. (2023). Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings. *Political Analysis*, *31*(3), 366–379. https://doi.org/10.1017/pan.2022.29

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. *Journal of Machine Learning Research*, *24*(253), 1–15.

Müller, P., Chan, C.-H., Ludwig, K., Freudenthaler, R., & Wessler, H. (2023). Differential Racism in the News: Using Semi-Supervised Machine Learning to Distinguish Explicit and Implicit Stigmatization of Ethnic and

Religious Groups in Journalistic Discourse. *Political Communication*, *40*(4), 396–414. https://doi.org/10.1080/10584609.2023.2193146

Palmer, A., Smith, N. A., & Spirling, A. (2024). Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, *4*(1), 2–3. https://doi.org/10.1038/s43588-023-00585-1

Plaza-del-arco, F. M., Nozza, D., & Hovy, D. (2023). Respectful or Toxic? Using Zero-Shot Learning with Language Models to Detect Hate Speech. *The 7th Workshop on Online Abuse and Harms (WOAH)*, 60–68. https://doi.org/10.18653/v1/2023.woah-1.6

Puschmann, C., Karakurt, H., Amlinger, C., Gess, N., & Nachtwey, O. (2022). RPC-Lex: A dictionary to measure German right-wing populist conspiracy discourse online. *Convergence*, *28*(4), 1144–1171. https://doi.org/10.1177/13548565221109440

Puschmann, C., Stier, S., Merten, L., Kulshrestha, J., Rauxloh, H., & Schultz, C. (2023). *German Online News Domains*. https://doi.org/10.17605/OSF.IO/S5UHB

Reiss, M. V. (2023). *Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark* (No. arXiv:2304.11085). arXiv. https://doi.org/10.48550/arXiv.2304.11085

Rillig, M. C., Ågerstrand, M., Bi, M., Gould, K. A., & Sauerland, U. (2023). Risks and Benefits of Large Language Models for the Environment. *Environmental Science & Technology*. https://doi.org/10.1021/acs.est.3c01106

Rogers, R., & Zhang, X. (2024). The Russia–Ukraine War in Chinese Social Media: LLM Analysis Yields a Bias Toward Neutrality. *Social Media + Society*, *10*(2), 20563051241254379. https://doi.org/10.1177/20563051241254379

Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., & Gadepally, V. (2023). *From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference* (No. arXiv:2310.03003). arXiv. https://doi.org/10.48550/arXiv.2310.03003

Soroka, S., Young, L., & Balmas, M. (2015). Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content. *The ANNALS of the American Academy of Political and Social Science*, *659*(May), 108–121. https://doi.org/10.1177/0002716215569217

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*, *616*(7957), 413–413. https://doi.org/10.1038/d41586-023-01295-4

Stoll, A. (2023). The accuracy trap or How to build a phony classifier. In *Challenges and perspectives of hate speech research* (pp. 371–381). https://doi.org/10.48541/dcr.v12.22

Tekumalla, R., & Banda, J. M. (2023). Leveraging Large Language Models and Weak Supervision for Social Media Data Annotation: An Evaluation Using COVID-19 Self-reported Vaccination Tweets. In H. Mori, Y. Asahi, A. Coman, S. Vasilache, & M. Rauterberg (Eds.), *HCI International 2023 – Late Breaking Papers* (pp. 356–366). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-48044-7_26

Theocharis, Y., & Jungherr, A. (2021). Computational Social Science and the Study of Political Communication. *Political Communication*, *38*(1–2), 1–22. https://doi.org/10.1080/10584609.2020.1833121

Törnberg, P. (2024a). *Best Practices for Text Annotation with Large Language Models* (No. arXiv:2402.05129). arXiv. https://doi.org/10.48550/arXiv.2402.05129

Törnberg, P. (2024b). Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages. *Social Science Computer Review*, 08944393241286471. https://doi.org/10.1177/08944393241286471

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). *Efficient Few-Shot Learning Without Prompts* (No. arXiv:2209.11055). arXiv. https://doi.org/10.48550/arXiv.2209.11055

van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N., & Schlobach, S. (2008). Good News or Bad News? Conducting Sentiment Analysis on Dutch Text to Distinguish Between Positive and Negative Relations. *Journal of Information Technology & Politics*, *5*(1), 73–94. https://doi.org/10.1080/19331680802154145

van Atteveldt, W., & Peng, T.-Q. (2018). When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Viehmann, C., Beck ,Tilman, Maurer ,Marcus, Quiring ,Oliver, & and Gurevych, I. (2023). Investigating Opinions on Public Policies in Digital Media: Setting up a Supervised Machine Learning Tool for Stance

Classification. *Communication Methods and Measures*, *17*(2), 150–184. https://doi.org/10.1080/19312458.2022.2151579

Windsor, L. C. (2021). Advancing Interdisciplinary Work in Computational Communication Science. *Political Communication*, *38*(1–2), 182–191. https://doi.org/10.1080/10584609.2020.1765915

Xu, H., Lou, R., Du, J., Mahzoon, V., Talebianaraki, E., Zhou, Z., Garrison, E., Vucetic, S., & Yin, W. (2024). *LLMs' Classification Performance is Overclaimed* (No. arXiv:2406.16203). arXiv. https://doi.org/10.48550/arXiv.2406.16203

Yang, K.-C., & Menczer, F. (2025). Accuracy and Political Bias of News Source Credibility Ratings by Large Language Models. *Proceedings of the 17th ACM Web Science Conference 2025*, 127–137. https://doi.org/10.1145/3717867.3717903

Zarouali, B., Araujo, T., Ohme, J., & de Vreese, C. (2024). Comparing Chatbots and Online Surveys for (Longitudinal) Data Collection: An Investigation of Response Characteristics, Data Quality, and User Evaluation. *Communication Methods and Measures*, *18*(1), 72–91. https://doi.org/10.1080/19312458.2022.2156489

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, *50*(1), 237–291. https://doi.org/10.1162/coli_a_00502

Appendix A: LLM prompt

Your task is to classify evaluations in a sample of German-language newspaper article extracts that relate to climate change.

The text shown below describe events surrounding climate change related protests or measurements, either protests demanding action or reports on a speedlimit in German traffic or the reformation of German heating infrastructure, such as the implementation of heat pumps and replacements of traditional heating systems.

Your task is structured as follows:

1) Label the text as either SUPPORTIVE of or OPPOSED to the protests or the measure in question, or as NEUTRAL.

A statement is SUPPORTIVE if it expresses support, sympathy, or understanding for the movement or its goals.

A supportive statement praises the benefits in the context of sustainability, fighting climate change, and improving social welfare.

Merely reporting decisions or juridical reformation without highlighting benefits does not qualify as an endorsement.

A statement is OPPOSED if it expresses criticism or antipathy towards the protest or measurement, or demonstrates a lack of understanding or principal opposition towards it.

Both SUPPORTIVE and OPPOSED statements could also be implicitly indicated or reflected by a quoted or referenced actor, not necessarily the author of the text that is presented.

If such SUPPORTIVE or OPPOSED positions are echoed by the text, still code this as a tendency towards either opposition or support.

2) If the evaluation is either SUPPORTIVE or OPPOSED, please rate the strength of the evaluation with a value between 1 (extremely weak support or extremely weak opposition) and 5 (extremely strong support or opposition). If the evaluation is NEUTRAL, return NA.

3) Also provide a short chain-of-thought-reasoning for you decision in an extra column.

Always respond in English. Always format your answer as strictly valid JSON with the keys "label", "strength", and "reasoning".