Clarissa Hohenwalde[1], Melanie Leidecker-Sandmann[1], Nikolai Promies[1],*
& Markus Lehmkuhl[1]

[1] Karlsruhe Institute of Technology, Department of Science Communication, Germany
[*] corresponding author: ai3551@kit.edu

ChatGPT’s Potential for Quantitative Content Analysis: Categorizing Actors in Public Debates

ABSTRACT:
We assess ChatGPT's ability to identify and categorize actors in news media articles into different societal groups. We conducted three experiments to evaluate different models and prompting strategies. In experiment 1, testing gpt-3.5-turbo, we found that using the original codebooks created for manual content analysis is insufficient. However, combining named entity recognition with an optimized prompt (NERC pipeline) yielded an acceptable macro-averaged F1-score of .79. Experiment 2 compared gpt-3.5-turbo, gpt-4o, and gpt-4-turbo: the latter achieved the highest macro-averaged F1-score of .82 using the NERC pipeline. Challenges remained in classifying nuanced actor categories. Experiment 3 demonstrated high retest reliability for different gpt-4o releases.

Keywords: ChatGPT, coding, categorizing, quantitative content analysis, automated content analysis, artificial intelligence, Large Language Models (LLMs), actors, AI tools in science communication

Introduction

Content analysis is a fundamental research method in communication science and the most widely used empirical technique in the field [Brosius et al., 2022; Gómez-Escalonilla, 2021; Nicolás et al., 2019; Riffe & Freitag, 1997; Trumbo, 2004]. Traditionally, it entails training human coders to classify texts through an iterative process based on detailed codebooks. A weakness of this process is that it is both time-consuming and financially costly. During the coding phase, the workload linearly depends on the number of units to be examined [Brosius et al., 2022], making large-scale studies and real-time analysis difficult to manage in manual analysis. This is increasingly problematic given the proliferation of diverse fragmented digital content in a digital world [Kroon et al., 2024].

To address these challenges, efforts have been made to further develop and refine content analysis through computer-aided and automated methods [Buz et al., 2022; Brosius et al., 2022; Haim, 2023; Scharkow, 2013; Wirth et al., 2015]. Compared to manual coding, automated coding relies primarily on computational resources, making big data analyzes more manageable, facilitating real-time analysis, and offering significant financial savings [e.g. Scharkow, 2013]. However, the application of automated content analysis methods is not yet part of the standard training in communication science and many automated methods require advanced programming skills, which poses a significant entry barrier to social science researchers [Strippel et al., 2018].

Unlike other advanced machine learning approaches, the recently emerged large language models (LLMs) like ChatGPT can be prompted using natural language. As powerful artificial intelligence models, LLMs utilize machine learning techniques to process human language and generate coherent text [Gill & Kaur, 2023]. Their flexibility and advanced natural language processing capabilities make them particularly interesting for content analysis tasks. Despite their potential, the application of LLMs in communication science, specifically for quantitative and qualitative content analysis, is still emerging. This presents an opportunity to explore their effectiveness in automating traditional content analysis tasks.

A typical task in content analysis is the identification and classification of actor groups within journalistic media articles, which this study focuses on. We aim to assess the potential of ChatGPT to replace human coders in quantitative content analysis, thereby contributing to the advancement of automated content analysis methods in communication science. In this study, we conducted three exploratory experiments to evaluate the performance of different prompting strategies and ChatGPT models for the identification and classification of actors.

This paper is organized as follows: First, we provide an overview of how automated content analysis and especially LLM-based approaches have been utilized in social and communication science research. We then elaborate on the significance of actor coding and classification from a communication science perspective. Subsequently, we present the methodological procedures of our three experimental analyses and describe their results. Finally, we offer a critical discussion of the findings, including limitations and implications for future research in journalism studies.

State of Research

Automated content analysis has a long-standing tradition in communication science, predating the recent development of large language models (LLMs). In the early 1960s, Stone et al. [1962] introduced the General Inquirer, a software tool that enabled automated coding of written text and spoken language based on dictionaries. Such approaches conduct frequency analyses by counting the occurrence of words or phrases to determine topic prevalence or classify texts [Brosius et al., 2022]. A popular application of dictionaries is sentiment analysis that captures positive and negative emotions expressed through text [Boumans & Trilling, 2016]. While valuable for variables that share repetitive characteristics [Günther & Quandt, 2016], these approaches have notable limitations: Their validity is closely tied to the context for which the dictionary is developed (reduced generalizability) and the inclusion or exclusion of specific terms is subject to researcher bias [Burscher et al., 2015; Kroon et al., 2024].

Over time, advancements in machine learning introduced supervised learning techniques to content analysis. These techniques commonly leverage bag-of-words models and have been shown to outperform dictionaries in various contexts [Burscher et al., 2015; Kroon et al., 2024; Scharkow, 2011]. In supervised methods, machine learning models are trained on manually labeled input data to inductively create statistical models that strive to replicate the manual coding results [Brosius et al., 2022; Scharkow, 2011]. This approach allows for better accommodation of textual data particularities and reduces researcher bias by letting the algorithm infer patterns that lead to specific classifications [Burscher et al., 2015; Chew et al., 2023; Kroon et al., 2024]. However, supervised machine learning requires substantial amounts of manually annotated training data to ensure validity, which especially poses a challenge for social science research, where new studies often necessitate domain-specific training data tailored to specific inquiries [Chew et al., 2023; Laurer et al., 2024; Törnberg, 2023b]. Moreover, these supervised methods do not generalize well across different languages, domains, or genres, further limiting their utility [Kroon et al., 2024].

To alleviate these issues, pretrained models have been introduced, providing foundational linguistic understanding that can be adapted to specific tasks [Brosius et al., 2022]. Named Entity Recognition (NER) is a widely used pretrained method that extracts information from unstructured texts to identify named entities such as people, organizations, or locations [Marrero et al., 2013, Schneider, 2014]. NER techniques are particularly relevant for identifying actors in texts, and studies have demonstrated their effectiveness in journalistic contexts [Buz et al., 2022]. However, adapting or fine-tuning pretrained models still requires considerable programming knowledge and does not entirely eliminate the need for large labeled datasets.

The emergence of LLMs however offers new possibilities: Models based on the transformer architecture are trained on vast amounts of unstructured text data, enabling them to obtain transferable language knowledge applicable to various downstream tasks without extensive task-specific training data [Brown et al., 2020, Kroon et al., 2024, Törnberg, 2023b]. This makes them particularly attractive for content analysis in communication science. Applications like ChatGPT allow researchers to use natural language prompts to perform tasks such as text classification. In a so-called zero-shot classification setting, the model assigns texts to new categories it was not explicitly trained on by inferring from class descriptors in the prompt [Brown et al., 2020]. Few-shot learning further improves generalization performance by supplying a small number of labeled examples [Brown et al., 2020]. All in all, prompting LLMs like ChatGPT 1) requires minimal technical expertise and is more accessible than traditional automated content analysis methods, 2) eliminates the need for large labeled datasets and 3) is relatively inexpensive and quick to implement. This makes it an appealing option for researchers seeking efficient analytical tools.

While scholars have suggested potential use cases for LLMs in social science research [Argyle et al., 2023; Binz & Schulz, 2023; Stokel-Walker & Van Noorden, 2023], their application remains limited. As automated methods gain prominence, there is an increasing need for methodological discussions about the quality requirements, validity and reliability of individual methods [Buz et al., 2022; Niekler, 2018], a gap that our study seeks to address.

Recent pioneering studies across various disciplines have examined ChatGPT's coding capabilities in direct comparison to human coders, revealing both strengths and limitations. Research by Gilardi et al. [2023] found that ChatGPT-3.5-turbo achieved a higher accuracy for most topic and frame classification tasks in news articles and tweets than Amazon Mechanical Turk crowdworkers compared to a baseline of trained coders. Similarly, Zambrano et al. [2023] evaluated ChatGPT-4’s ability to identify socially positive and negative constructs in press related texts, showing that while the model performed well with clear categories, it struggled with more ambiguous tasks, leading to a tendency to overgeneralize. Further studies indicate that ChatGPT-4 can accurately infer political affiliations from social media content, often outpacing human coders [Törnberg, 2023a]. Compared to a fine-tuned transformer-based language model, ChatGPT-3.5 performs better when categorizing texts into genre categories in a zero-shot setting [Kuzman et al., 2023]. However, Xiao et al. [2023] and Tai et al. [2024] emphasize that the effectiveness of ChatGPT-3.0 and ChatGPT-3.5 is highly influenced by prompt design. Prompts providing a clear codebook and examples for coding yielded results comparable to manual coding [Xiao et al., 2023]. By offering contextual information and structure, following the form “code - description - examples”, the coding proved to be more reliable than results archived by pure example-centered prompts without additional context [Xiao et al., 2023]. Tai et al. [2024] note that conducting multiple iterations of coding with ChatGPT aids in obtaining more consistent results.

With regards to using ChatGPT to automatically classify actors in texts, which is the focus of our study, we are only aware of a recent, yet unpublished study by Wiesner [2024] that successfully employed GPT4-o to identify references to scientific actors (individual and institutional) in more than 230,000 written parliamentary speeches in the Austrian Nationalrat. The speeches were firstly divided into sequences identified by a dictionary approach and secondly analyzed by ChatGPT using a simple prompting strategy (roughly: Is there a reference to a scientist or a scientific institution in the text?). However, Wiesner [2024] - to our knowledge - did not further differentiate among types of scientists, identify other actor groups, test different GPT models, or systematically evaluate the automated coding.

Overall, a literature review by Ollion et al. [2023] shows that while LLMs like ChatGPT often match human performance when analyzing texts, their effectiveness in coding tasks is partial and varies based on material, language, and prompt. This variability underscores the necessity for further research to understand and enhance the capabilities of LLMs in content analysis tasks.

Why Actor Classification as a Test Case?

There are potentially many content analytic tasks that we could have chosen as a test case for coding with ChatGPT. However, we argue that actor identification and classification within journalistic media articles serves as an ideal test case for exploring ChatGPT's potential in automated content analysis.

While this task requires an advanced and nuanced understanding of language and context from a purely technical perspective, analyses of the structure of actors in public discourse are also highly relevant to communication science and interesting from a sociological perspective. The public sphere in modern societies can be seen as a forum or arena for public communication and the exchange of opinions [Gerhards & Neidhardt, 1990; Habermas, 1992]. “Because the public sphere is centrally located in the forecourt of power in the topography of society, it is always a contested area. Actors in society try to assert their issues and make their opinions plausible as generalizable opinions.” [Gerhards & Neidhardt, 1990, p. 11, translation by the authors].

Assessing which actors are granted access to the public “arena” is relevant because actors from various societal groups strive to promote their perspectives on important issues, thereby influencing public opinion and policymaking [Gerhards & Neidhardt, 1990, p. 27]. From the perspective of deliberative theories of democracy and the public sphere it seems (normatively speaking) important that representatives of *all* social groups can participate in the public discourse in order to ensure the free formation of opinion among the members of society (recipients in the “gallery” [Gerhards & Neidhardt, 1990, p. 27] - in particular, when socially relevant problems are discussed publicly [Habermas, 1992]. Only the voices of actors represented in public media discourse become visible or audible, impacting the collective shaping of public opinion. An example is the issue of retail closures during the COVID-19 pandemic: It makes a difference whether this issue is discussed predominantly with reference to economic actors, or from the perspective of predominantly scientific actors. The “carrying capacity” [Gerhards & Neidhardt, 1990, p. 27] of the media in terms of actors in the public sphere is limited, leading to competition among actors for participation in media discourses.

Many empirical studies deal with the diversity of actors in media coverage, examining when and how often specific actors are mentioned [e.g. Albæk et al., 2003; Leidecker-Sandmann & Lehmkuhl, 2022a; Leidecker-Sandmann et al., 2022b; Burggraaff & Trilling, 2020; Eisenegger et al., 2020; Maurer et al., 2021; Niekler, 2018]. These studies enable researchers to make statements about the visibility and relevance of certain actors or groups within specific public debates, contributing to a deeper understanding of media influence on societal issues.

Given the importance of actor analysis in communication science and the challenges associated with manual coding, actor identification and classification presents a meaningful and practical test case for evaluating the effectiveness of ChatGPT in automated content analysis. By focusing on this task, we aim to assess whether ChatGPT can reliably identify and categorize societal actors in journalistic texts, thereby offering a scalable and efficient alternative to traditional manual coding methods.

Methods

To evaluate the potential of ChatGPT in replacing human coders for quantitative content analysis, we conducted three exploratory experiments. We tested various prompting strategies and model versions, assessing their performance against manual coding in terms of precision, recall, overall accuracy, and reliability.

Sample Description

As a case study, we chose the quantitative content analysis task of identifying and categorizing actor groups into different societal groups in science-related media debates. For this purpose, we analyzed German print media coverage on four key scientific topics — biotechnology, climate change, neuroscience, and antibiotic resistance — where significant reporting and the involvement of scientists were expected. The media sample encompassed both mainstream and regional media outlets within the German media landscape. It included national news magazines (Der Spiegel, Der Stern), national daily newspapers (Die Welt, taz), and regional newspapers (Berliner Zeitung, Nürnberger Nachrichten). To capture temporal variations in media coverage, two distinct investigation periods were set for each issue, except for antibiotic resistance, which was sampled continuously due to fewer articles. The time frames with significant media reporting were selected pragmatically. Articles were accessed through the Nexis Uni database [Nexis Uni, 2024] using specific search strings tailored to each issue (see Appendix), resulting in a total sample of 2,883 articles (see Table 1).

Issue	Investigation period	Number of articles (initial manual coding)
Biotechnology	01.01.2000 — 31.12.2001 + 01.01.2016 — 31.12.2019	810
Climate change	01.01.2000 — 31.12.2001 + 01.01.2018 — 31.12.2019	891
Neuroscience	01.01.2000 — 31.12.2001 + 01.01.2017 — 31.12.2019	612
Antibiotic resistance	Articles for each year from 01.01.2000 — 31.12.2019	570
		Σ 2,883

Table 1: Investigation periods and number of articles by issue.

Baseline

To establish a baseline against which we could compare ChatGPT's coding performance, we first conducted a semi-automatic content analysis to identify and classify actors mentioned in the articles into their respective societal groups.

To assist in detecting potential actors within the texts, we employed an automated Named Entity Recognition (NER) approach. The NER tool automatically extracted named entities (individual persons) from the articles. The NER process was conducted using the Python-based package FLAIR [Akbik et al., 2019], which has demonstrated high precision and recall in previous validations for German journalistic texts, having extracted 99 % of relevant individual actors with minimal irrelevant or incomplete results in a prior analysis [Buz et al., 2022]. This high level of accuracy provided a robust foundation for subsequent actor classification tasks.

An elaborate coding scheme was developed to classify the pre-identified actors into societal groups based on their roles and affiliations (see Appendix). Each actor was assigned to only one category based on their primary role or affiliation; multiple categorizations were not permitted to ensure clarity and consistency in the coding process. Following Habermas [1992] and an aggregation of the various social positions of the political system according to Easton [1990], we distinguish between actors in the public decision-making process who, ideally, can be assigned either to the so-called 'center' (executive, legislature, parties, political administration) or the so-called 'periphery'. The latter includes, among others, scientific actors, but also organized associations, trade unions or other associations of social particular interests as well as so-called public interest groups that represent collective goods interests (e.g. the environment, animals, consumers). We therefore distinguish between the following categories of actors in our analysis:

- Researchers: Individuals conducting scientific research without political or social functions.
- Science Administration: Personnel involved in research policy and administration, including those at federal research institutions and international organizations such as the US-American CDC (Centers for Disease Control) or the World Health Organization (WHO).
- Medical Experts: Practicing physicians and medical professionals.
- Politicians: Members of executive functions, administrative bodies, and legislative assemblies.
- Advocacy Groups: Representatives of collective or partial interests, such as non-governmental organizations (NGOs) and lobbyists.
- Other Actors: Individuals from peripheral societal domains in the Habermasian sense.

A team of 20 coders was assembled to classify the actors identified by the NER process. The coders underwent multiple training sessions and were provided with a detailed coding manual as well as additional example codings for each category to ensure a consistent understanding of the classification scheme.

To assess inter-coder reliability, all coders independently coded a subset of 13 articles (3 to 4 articles per issue), encompassing a total of 163 actor mentions. The results were compared against a master coding established by an expert coder (the study director). The Krippendorff's alpha calculated for all categories combined was α = .63, indicating a rather low level of agreement [Krippendorff, 2018]. This highlighted challenges in differentiating between closely related categories such as “Researchers”, “Science Administration”, and “Medical Experts”. Due to the observed difficulties, the coding scheme was post-hoc simplified by consolidating the mentioned categories into a single category labeled "Science". This adjustment raised the overall Krippendorff's alpha to α = .77, which is considered acceptable for tentative conclusions in content analysis [Krippendorff, 2018].

The moderate agreement levels can be attributed to the nuanced distinctions between actor categories. For instance, the boundaries between politics and science can become blurred, especially in domains involving federal research institutions where roles may overlap. Additionally, actors representing partial interests, such as researchers working for private companies (e.g. Bayer or BASF), were sometimes misclassified, underscoring the complexity of accurately categorizing actors based solely on textual mentions.

Following the reliability assessment and coding scheme adjustment, the coders proceeded to classify the actors in the remaining articles, resulting in a comprehensive dataset of actor classifications that serves as a basis of comparison for the experiments with ChatGPT conducted and presented in this paper.

Experiment 1: Evaluating Prompting Strategies

The first experiment focused on determining whether ChatGPT could identify and code actors with high validity using different prompting strategies. At the time of experiment 1, gpt-3.5-turbo was the latest model available from OpenAI. We explored three approaches to prompt the model:

1. Zero-shot prompting with a detailed codebook: We prompted ChatGPT using the detailed codebook initially designed for human research assistants. The codebook contained exhaustive definitions for each actor category. This approach aims to assess whether codebooks have to be adjusted for this type of automatic quantitative content analysis and whether prior examples or additional context are needed.

2. Few-shot learning with an optimized prompt: Recognizing the potential limitations of zero-shot prompting, we optimized the prompt by applying few-shot learning principles. Instead of providing exhaustive definitions, we supplied the model with category keywords and illustrative examples. This approach aimed to enhance the model's understanding by providing minimal guidance and leveraging its ability to generalize.

3. Integration into a NERC pipeline: In the third approach, we integrated ChatGPT into a Named Entity Recognition and Classification (NERC) pipeline to reduce task complexity. We performed automatic Named Entity Recognition (NER) using the FLAIR package in Python [Akbik et al., 2019], which was also used for the initial human coding. This meant that ChatGPT only needed to classify the identified actors. The extracted names, along with small text windows containing these names for contextualization, were then passed to ChatGPT using the optimized prompt from the second approach. This method provided the model with both the entity and its immediate context within the article, enhancing its ability to accurately classify the actors.

The full prompts can be found in the Appendix.

For prompt development and optimization, we used a subsample of 100 articles from the previously described dataset. The articles were equally distributed across the four scientific issues, ensuring a balanced representation of topics. This subsample allowed us to refine our prompts without risking overfitting.

To evaluate ChatGPT's performance in a quantitative analysis setting, we utilized a distinct sample of 200 articles, again equally distributed across the four scientific issues for testing the prompt strategies.

All interactions with ChatGPT were conducted with a temperature setting of .0 and a top_p value of .5 to ensure deterministic and consistent outputs. We compared ChatGPT's classifications against our human-coded dataset, which served as the gold standard[1]. Given the multi-class classification problem with imbalanced class distribution, we evaluated the overall performance by using a macro-averaged F1-score. This approach balances precision and recall across all classes, treating each class equally regardless of its size. To further substantiate findings in cases where the results offered valuable insights for strengthening the overall argument, we analyzed class F1-scores, class precision, class recall and the confusion matrix. This allowed us to identify categories with frequent misclassifications and to pinpoint areas where category distinctions required further refinement.

Experiment 2: Comparing Different GPT Models

The second experiment investigated whether the underlying GPT model influenced classification outcomes. We compared the performance of gpt-3.5-turbo with two subsequently released models: gpt-4-turbo and gpt-4o.

- GPT-3.5-Turbo: Introduced on March 15, 2022, this model served as our baseline.
- GPT-4-Turbo: Released on November 6, 2023, gpt-4-turbo can process the equivalent of more than 300 pages of text in a single prompt, has a broader training knowledge up to December 2023 and is better than previous models at carefully following instructions [OpenAI, 2023].
- GPT-4o: Unveiled on May 13, 2024, gpt-4o represents a significant advancement toward natural human-computer interaction. It accepts any combination of text, audio, and image inputs and produces outputs in text, audio, or image formats. The performance of gpt-4o matches that of gpt-4-turbo on English texts and code but offers substantial improvements in processing non-English languages, including German. Additionally, the gpt-4o API provides faster response times and is 50 % more cost-effective [OpenAI, 2024].

Using the test sample of 200 articles from experiment 1, we applied the same three prompting strategies as before. We maintained the temperature at 0 and top_p at .5 to ensure consistency across models. The performance of each model was again evaluated against the human-coded gold standard using the macro-averaged F1-score and F1-scores for each category. To identify any systematic biases or patterns in misclassification among the different models, we analyzed class F1-scores, class precision, class recall and the confusion matrix for selected cases.

[1] In this case, the term “gold standard” is not to be understood in the sense of perfectly reliable coding, but rather in the sense of a basis for comparison, which itself has weaknesses (as already described).

Experiment 3: Assessing Model Reliability Across Releases

The third experiment aimed to assess the retest reliability of ChatGPT's output and its stability across different releases of the same model. Given that OpenAI continually updates its models, it is crucial to determine whether the model produces consistent classifications over multiple requests and whether updates affect performance.

Due to economic reasons, we focused on the cheaper gpt-4o model, examining versions from May 13 and August 6. Using the same sample of 200 articles, we employed the NERC pipeline for coding, maintaining the temperature at 0 and top_p at .5. Each version of gpt-4o was used to code all actors from the articles ten times, allowing us to assess both intra-release and inter-release reliability.

To evaluate reliability, we calculated Krippendorff's alpha for each gpt-4o version individually and for both versions combined. To obtain confidence intervals for the Krippendorff's alpha values, we performed bootstrapping with 1,000 iterations for each dataset. This approach provided robust estimates of reliability and allowed us to assess the stability of ChatGPT's classifications over time.

Statistical Analysis

Statistical analyses were performed using a combination of Python and R. For classification performance, we calculated precision, recall, and F1-scores for each category and generated confusion matrices to visualize misclassifications and identify patterns of errors using Python. Krippendorff's alpha was computed in R version 4.3.1 using the irr package, with bootstrapping performed using the boot package to estimate confidence intervals [Canty & Ripley, 2024; Gamer & Lemon, 2019; R Core Team, 2023].

In the evaluation of the models, we paid special attention to the imbalanced nature of the class distribution. We employed macro-averaged F1-scores to ensure that performance metrics were not unduly influenced by the majority classes.

Results

Experiment 1: Evaluating Prompting Strategies

In experiment 1, we assessed whether ChatGPT could be effectively prompted to produce valid results for the identification and classification of actors according to their societal roles. We focused exclusively on the gpt-3.5-turbo model and tested three prompting strategies: 1) using the original codebook designed for human coders (zero-shot prompting), 2) employing an optimized prompt utilizing few-shot learning principles, and 3) integrating ChatGPT into a Named Entity Recognition and Classification (NERC) pipeline.

Prompt	Codebook			Optimized prompt			NERC pipeline		
Class	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Science	.35	.79	.49	.38	.75	.50	.91	.96	.94
Advocacy	.00	.00	.00	.32	.55	.41	.84	.70	.76
Politics	.35	.69	.47	.34	.55	.42	.92	.91	.91
Others	.05	.28	.08	.05	.27	.09	.57	.55	.56
Overall	**.19**	**.44**	**.26**	**.27**	**.53**	**.36**	**.81**	**.78**	**.79**

Table 2: Precision, Recall and F1-scores for different prompting strategies using gpt-3.5-turbo.

When prompted using the detailed codebook intended for human coders (zero-shot prompting), gpt-3.5-turbo achieved an overall F1-score of .26 (see Table 2). The model correctly classified some actors in the "Science" and "Politics" categories but struggled significantly with accurately identifying actors in the "Advocacy" and "Other Actors" categories. This resulted in low precision and recall for these categories.

		PREDICTED				
		Science	**Advocacy**	**Politics**	**Others**	∑
EXPECTED	**Science**	182	3	0	3	188
	Advocacy	7	37	0	3	47
	Politics	5	3	41	1	50
	Others	7	2	0	8	17
	∑	201	45	41	15	302

Table 3: Confusion matrix for gpt-3.5-turbo using an optimized prompt.

Employing the optimized prompt with few-shot learning principles led to a modest improvement in classification performance. The overall F1-score increased to .36. This approach enhanced the classification of "Advocacy" actors, as reflected in higher F1-scores for that category. Misclassifications nevertheless occurred and often involved advocacy actors being incorrectly classified as "Science" actors (see Table 3). For instance, in the following example Craig Venter is classified as "Science", despite representing partial interests as the CEO of a gene technology company.

"Last Thursday, US researcher Craig Venter announced that his genetic engineering company 'Celera Genomics' had decoded (sequenced) 99 percent of the entire human genome [...]"[2].

		PREDICTED				
		Science	**Advocacy**	**Politics**	**Others**	∑
EXPECTED	**Science**	234	3	1	3	241
	Advocacy	12	47	0	8	67
	Politics	3	3	67	1	74
	Others	8	5	0	16	29
	∑	257	58	68	28	421

Table 4: Confusion matrix for the integration of ChatGPT into the NERC pipeline.[3]

The highest performance was achieved when integrating ChatGPT into the NERC pipeline. By first performing automatic NER to extract named entities and then providing ChatGPT with these entities along with contextual text snippets from the articles, the model's overall F1-score improved significantly to .79. Precision and recall values increased across all categories, indicating that providing context and focusing on identified entities enhanced the model's ability to classify actors accurately. Analysis of the confusion matrices revealed that, even with the NERC pipeline, gpt-3.5-turbo encountered difficulties with the "Other Actors" category (see Table 4). The model's precision and recall for this category were .57 and .55 respectively, suggesting frequent misclassifications. One example for this is the following snippet: “Company boss Detlev Goj sends out an average of 400 to 500 parcels of sterile maggots every month”[4]. Here, the CEO Detlev Goj was classified as “Other Actors”, even though he represents his company and should therefore be classified as “Advocacy”.

Experiment 2: Comparing Different GPT Models

In the second experiment, we evaluated the performance of three GPT models (gpt-3.5-turbo, gpt-4-turbo and gpt-4o) across the three prompting strategies previously

[2] Translation by the authors. Original text: "Da verkündete der US-Forscher Craig Venter am vergangenen Donnerstag, seine Gentechnikfirma 'Celera Genomics' habe das gesamte menschliche Erbgut zu 99 Prozent entschlüsselt (sequenziert)”

[3] The number of actors differs from Table 3, as more actors were identified by NER than by ChatGPT.

[4] Translation by the authors. Original text: “Durchschnittlich 400 bis 500 Pakete mit sterilen Maden versendet Firmenchef Detlev Goj monatlich.”

tested: using the original codebook designed for human coders, employing an optimized prompt with few-shot learning principles, and integrating ChatGPT into a named entity recognition and classification (NERC) pipeline. The objective was to determine whether advancements in model architecture and capabilities influenced classification outcomes for the given task.

Prompt	Codebook	Optimized prompt	NERC pipeline
Model			
gpt-3.5-turbo	.26	.36	.79
gpt-4-turbo	.35	.39	.82
gpt-4o	.38	.42	.70

Table 5: Comparison of F1-scores for different prompting strategies and different models.

When prompted with the original codebook intended for human researchers in a zero-shot setting, all three models demonstrated relatively low overall F1-scores (see Table 5). While some actors in the "Science" and "Politics" categories were correctly identified and classified by all models, there was a consistent struggle in accurately identifying and classifying actors in the "Advocacy" and "Other Actors" categories. The low F1-scores indicate that merely providing a detailed codebook without additional context or examples was insufficient for high-quality classification across models.

Using the optimized prompt that included category keywords and illustrative examples led to modest improvements in performance. This approach primarily enhanced the identification and classification of "Advocacy" actors across models. However, the overall improvement was limited, and misclassifications persisted.

Integrating ChatGPT into the NERC pipeline yielded the highest F1-scores across all models. With the NERC pipeline, gpt-3.5-turbo showed substantial improvements but still encountered difficulties with certain categories as described in experiment 1. Gpt-4-turbo demonstrated the highest overall F1-score among the models when using the NERC pipeline. The model showed improved performance in the "Other Actors" category, with a precision of .75 and a recall of .52. Despite these improvements, misclassifications still occurred, particularly with actors from the "Other Actors" category being misclassified as "Science" or "Advocacy" actors. While gpt-4o achieved an almost perfect precision of .96 for the "Science" category, it struggled with the "Other Actors" category, which had a low precision of .23. Misclassifications were primarily due to "Science" and “Advocacy” actors being incorrectly coded as "Other Actors".

Experiment 3: Assessing Model Reliability Across Releases

For experiment 3, Krippendorff's alpha was calculated to evaluate the reliability of gpt-4o across different releases and over time. For the May 13 version, the alpha was α = .97 with a 95 % confidence interval of .95 to .98. The August 6 version showed an alpha of α = .98 with a 95 % confidence interval of .96 to .99. When combining the results from both releases, the overall Krippendorff's alpha was α = .98 with a 95 % confidence interval of .96 to .98. These high alpha values indicate consistent reliability both within each version and across versions over time.

Discussion

This study evaluated the potential of ChatGPT to replace human coders in the quantitative content analysis task of identifying and categorizing actor groups within German news media articles on biotechnology, climate change, neuroscience, and antibiotic resistance. Through three experiments, we assessed the validity, performance, and reliability of different ChatGPT models and prompting strategies.

Experiment 1: Evaluating Prompting Strategies

Experiment 1 demonstrated that using the original codebook designed for human coders was insufficient for achieving valid actor recognition and classification. Employing the original codebook with gpt-3.5-turbo yielded a low F1-score of .26. Optimized prompts improved the F1-score to .36; however, this value remains inadequate for reliable automatic coding.

The categories "Advocacy" and "Other Actors" posed significant challenges. We found that advocacy actors were frequently misclassified as "Science," primarily because scientists affiliated with businesses were categorized as "Science," despite the codebook considering them influenced by partial interests. This misclassification may be due to the overlapping characteristics of actors or insufficient contextual cues within the text. Actors representing partial interests or those with dual roles (e.g., a scientist owning a biotech company) present challenges for accurate categorization. Moreover, the "Other Actors" category, comprising actors not assignable to any specific group, proved difficult to classify accurately, as these actors are not semantically related. Addressing this issue may require a more nuanced categorization system to improve model accuracy.

The integration of ChatGPT into a Named Entity Recognition and Classification (NERC) pipeline substantially enhanced performance, improving the F1-score to .79, an acceptable value for coding purposes. This indicates that providing the model with pre-identified entities and contextual information is crucial for accurate classification. By isolating relevant entities and supplying contextual snippets, the model can better interpret the nuanced roles of actors within the text. This approach aligns with known prompting strategies that involve splitting complex tasks into subtasks, e.g. prompt chaining [Saravia, 2022].

Overall, our findings align with the observations of Tai et al. [2024] and Xiao et al. [2023] that the design of the prompt significantly influences coding results. However, contrary to Xiao et al. [2023], we found no advantage in supplying the codebook within the prompt.

Experiment 2: Comparing Different GPT Models

Experiment 2 compared the performance of gpt-3.5-turbo, gpt-4-turbo, and gpt-4o across the three prompting strategies. The NERC pipeline again yielded the highest F1-scores for all models, with gpt-4-turbo achieving the highest overall F1-score of F1 = .82. The superior performance of gpt-4-turbo over gpt-3.5-turbo highlights advancements in model capabilities, such as an expanded vocabulary and enhanced ability to recognize linguistic nuances. However, the fact that gpt-4o did not outperform gpt-4-turbo — despite its enhancements for non-English languages like German — suggests that model improvements do not necessarily translate linearly to better performance across all tasks.

Experiment 3: Assessing Model Reliability Across Releases

Experiment 3 assessed the reliability of gpt-4o by conducting multiple measurements on two different dates on which updates of the gpt-4o version were released. Krippendorff's alpha values were exceptionally high for both the May 13 (α = .97) and August 6 (α = .98) release of gpt-4o, with overlapping 95 % confidence intervals. The combined analysis yielded an alpha of α = .98, indicating consistent reliability both within and across model releases. The high reliability of gpt-4o over time suggests that ChatGPT can be used as a dependable tool for longitudinal studies. Consistent outputs across different releases alleviate concerns about the impact of updates of one and the same model on research reproducibility.

Limitations

Several limitations of this study should be acknowledged:

- The research focused exclusively on German-language media articles and four science related-media debates, which may limit the generalizability of the findings to other languages, topics, or text types.
- The actor classification scheme was simplified due to low inter-coder reliability among human coders. It is important to recognize that ChatGPT's coding is benchmarked against manual coding treated as the gold standard; however, this standard itself is not error-free, with an intercoder reliability of Krippendorff's alpha α = .77, indicating moderate human agreement.
- The imbalanced class distribution poses challenges for both human and automated classification. While we attempted to mitigate this by calculating macro-averaged F1-scores, the low number of actors in certain groups could bias performance metrics.

- We tested only one company's models, ChatGPT by OpenAI, which, despite achieving top-tier results in benchmarking large language models, may not be the most suitable for this specific task [LMaRena, 2024]. Models by other providers might offer better performance or different capabilities.
- OpenAI's models are proprietary, meaning that availability and cost are controlled by the company, potentially limiting the replicability of studies conducted with ChatGPT.
- Our reliance on pretrained models without domain-specific fine-tuning may limit the models' ability to capture specialized knowledge required for accurate classification.

Conclusions

This study demonstrates that ChatGPT holds considerable potential for automating actor classification in quantitative content analysis, particularly in the context of science-related news media articles. By integrating gpt-4-turbo into a named entity recognition and classification (NERC) pipeline and employing optimized prompting strategies, we achieved a valid coding outcome. This could significantly reduce the time and resources required for large-scale studies or media monitoring.

Our findings indicate that simply using codebooks designed for human coders is insufficient to achieve valid results when conducting quantitative content analysis with ChatGPT. To enhance performance, we recommend that researchers employ specific prompting strategies:

- Task Decomposition: Splitting complex tasks into subtasks, as implemented in our pipeline, allows the large language models to focus on manageable units, improving overall accuracy.
- Few-Shot Learning: Providing examples instead of only category definitions helps the model generalize and better understand nuanced distinctions between categories.
- Contextual Data Segmentation: Slicing data into appropriate text windows ensures that the model receives relevant contextual information, aiding in accurate classification.

Despite these improvements, persistent challenges remain in accurately classifying some items, particularly when categories are slightly overlapping or distinctions nuanced. This highlights the need for caution when employing ChatGPT for such quantitative content analysis, especially in contexts requiring precise distinctions.

Our experiments also revealed that newer models do not necessarily guarantee better performance for every task. While gpt-4-turbo achieved the highest macro-averaged F1-score in our study, gpt-4o did not outperform it despite enhancements for non-

English languages [OpenAI, 2024] Therefore, model selection should be guided by task-specific evaluations rather than assumptions based on model updates.

Notably, the high reliability exhibited by ChatGPT supports its feasibility for longitudinal research where consistency is paramount. The consistent outputs across different releases of the same model alleviate concerns about the impact of model updates on research reproducibility.

For future studies, we recommend:

- Validating Prompting Strategies: Utilize human-coded data to assess the effectiveness of your prompting approach by calculating F1-scores before deploying ChatGPT for automatic coding.
- Enhancing Classification Accuracy: Explore advanced prompting techniques, fine-tune language models on domain-specific corpora, or incorporate additional contextual information to improve performance in challenging categories.
- Assessing Generalizability: Expand analyses to include other languages, topics and text genres to determine the broader applicability of ChatGPT in content analysis tasks.
- Integrating with Other Tools: Investigate the combination of ChatGPT with other natural language processing tools to further enhance performance.

We believe that ChatGPT holds significant potential as a valuable asset for researchers in journalism studies and related fields, offering substantial advantages in efficiency and scalability. It can independently perform actor classifications or assist human coders as a collaborative tool, thereby enhancing the effectiveness and speed of content analysis tasks.

References

Akbik, A., Bergmann, T., & Vollgraf, R. (2019). *Flair: An easy-to-use framework for state-of-the-art NLP* (Version 0.11.3) [Computer software]. Zenodo. https://github.com/flairNLP/flair

Albæk, E., Christiansen, P. M., & Togeby, L. (2003). Experts in the mass media: Researchers as sources in Danish daily newspapers, 1961–2001. *Journalism & Mass Communication Quarterly*, 80(4), 937–948. https://doi.org/10.1177/107769900308000412

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

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120. https://doi.org/10.1073/pnas.2218523120

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

Brosius, H. B., Haas, A., & Unkel, J. (2022). Inhaltsanalyse III: Automatisierte Inhaltsanalyse. In Methoden der empirischen Kommunikationsforschung: Eine Einführung (pp. 179-194). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-34195-4

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., 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. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165

Burggraaff, C., & Trilling, D. (2020). Through a different gate: An automated content analysis of how online news and print news differ. *Journalism*, 21(1), 112-129. https://doi.org/10.1177/1464884917716699

Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using Supervised Machine Learning to Code Policy Issues: Can Classifiers Generalize across Contexts? The ANNALS of the American Academy of Political and Social Science, 659(1), 122-131. https://doi.org/10.1177/0002716215569441

Buz, C., Promies, N., Kohler, S., & Lehmkuhl, M. (2022). Validierung von NER-Verfahren zur automatisierten Identifikation von Akteuren in deutschsprachigen journalistischen Texten. *SCM Studies in Communication and Media*, 10(4), 590-627. https://doi.org/10.5771/2192-4007-2021-4-590

Canty, A. & Ripley, B. (2024). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-31. https://doi.org/10.32614/CRAN.package.boot

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., & Kim, A. (2023). LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. arXiv: 2306.14924. https://doi.org/10.48550/arXiv.2306.14924

Easton, D. (1990). The analysis of political structure. Routledge. https://doi.org/10.4324/9781003545798

Eisenegger, M., Oehmer, F., Udris, L., & Vogler, D. (2020). *Die Qualität der Medienberichterstattung zur Corona-Pandemie* (Qualität der Medien 1/2020). Forschungszentrum Öffentlichkeit und Gesellschaft (fög). http://www.foeg.uzh.ch/dam/jcr:ad278037-fa75-4eea-a674-7e5ae5ad9c78/Studie_01_2020.pdf

Gamer, M.. & Lemon, J. (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. https://cran.r-project.org/web/packages/irr/irr.pdf

Gerhards, J., & Neidhardt, F. (1990). *Strukturen und Funktionen moderner Öffentlichkeit. Fragestellungen und Ansätze*. WZB Discussion Paper No.

FS III 90-101. Berlin: Wissenschaftszentrum Berlin für Sozialforschung. https://bibliothek.wzb.eu/pdf/1990/iii90-101.pdf

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. *arXiv preprint* arXiv:2303.15056. https://doi.org/10.1073/pnas.2305016120

Gill, S. S., & Kaur, R. (2023). ChatGPT: Vision and challenges. Internet of Things and Cyber-Physical Systems, 3, 262-271. https://doi.org/10.1016/j.iotcps.2023.05.004

Gómez-Escalonilla, G. (2021). Métodos y técnicas de investigación utilizados en los estudios sobre comunicación en España. Revista mediterránea de comunicación, 12(1), 115-127. https://doi.org/10.14198/MEDCOM000018

Günther, E., & Quandt, T. (2016). Word counts and topic models: Automated text analysis methods for digital journalism research. *Digital Journalism*, 4(1), 75–88. https://doi.org/10.1080/21670811.2015.1093270

Habermas, J. (1992). *Faktizität und Geltung: Beiträge zur Diskurstheorie des Rechts und des demokratischen Rechtsstaats*. Frankfurt a. M.: Suhrkamp. https://doi.org/10.1007/978-3-531-90400-9_39

Haim, M. (2023). Texte als Daten I. In: Haim, M. (Ed.), *Computational Communication Science. Studienbücher zur Kommunikations- und Medienwissenschaft* (pp. 169-194). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-40171-9_8

Krippendorff, K. (2018). Content analysis: An introduction to its methodology. Sage publications. https://doi.org/10.4135/9781071878781

Kroon, A., Welbers, K., Trilling, D., & 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. *arXiv preprint* arXiv:2303.03953. https://doi.org/10.48550/arXiv.2303.03953

Laurer, M., Van Atteveldt, W., 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

Leidecker-Sandmann, M., & Lehmkuhl, M. (2022). Politisierung oder Aufklärung? Analysen der Akteur:innen- und Aussagenstruktur in medialen Diskursen über gesundheitliche Risikophänomene und die Rolle wissenschaftlicher Expert:innen. *Studies in Communication | Media*, 11(3), 337-386. https://doi.org/10.5771/2192-4007-2022-3-337

Leidecker-Sandmann, M., Attar, P., Schütz, A., & Lehmkuhl, M. (2022). Selected by expertise? Scientific experts in German news coverage of COVID-19 compared to other pandemics. *Public Understanding of Science*, 31(7), 847-866. https://doi.org/10.1177/096366252210957

LMaRena. (2024). Leaderboard. https://lmarena.ai/?leaderboard

Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., & Gómez-Berbís, J. M. (2013). Named entity recognition: Fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5), 482–489. https://doi.org/10.1016/j.csi.2012.09.004

Maurer, M., Reinemann, C., & Kruschinski, S. (2021). *Einseitig, unkritisch, regierungsnah? Eine empirische Studie zur Qualität der journalistischen Berichterstattung über die Corona-Pandemie*. https://rudolf-augsteinstiftung.de/wp-content/uploads/2021/11/Studie-einseitig-unkritisch-regierungsnahreinemann-rudolf-augstein-stiftung.pdf

Nexis Uni. (2024). News articles database. https://www.nexisuni.com

Nicolás, M. M., Saperas, E., & Carrasco-Campos, Á. (2019). La investigación sobre comunicación en España en los últimos 25 años (1990-2014). Objetos de estudio y métodos aplicados en los trabajos publicados en revistas españolas especializadas. Empiria. Revista de Metodología de las Ciencias Sociales, (42), 37-69. https://doi.org/10.5944/empiria.42.2019.23250

Niekler, A. (2018). *Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen*. Köln: Halem Verlag.https://doi.org/10.13140/RG.2.2.28090.39366

Ollion, E., Shen, R., Macanovic, A., & Chatelain, A. (2023). ChatGPT for Text Annotation? Mind the Hype. SocArXiv preprint. https://doi.org/10.31235/osf.io/x58kn

OpenAI. (2023, November 6). New models and developer products announced at DevDay. https://openai.com/index/new-models-and-developer-products-announced-at-devday/

OpenAI. (2024, May 13). Hello GPT-4o. https://openai.com/index/hello-gpt-4o/

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Riffe, D., & Freitag, A. (1997). A content analysis of content analyses: Twenty-five years of Journalism Quarterly. Journalism & Mass Communication Quarterly, 74(3), 515-524. https://doi.org/10.1177/107769909707400306

Saravia, E. (2022). Prompt engineering guide. GitHub. https://github.com/dair-ai/Prompt-Engineering-Guide

Scharkow, M (2011). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. Qual Quant 47, 761–773 (2013). https://doi.org/10.1007/s11135-011-9545-7

Scharkow, M. (2013). Automatische Inhaltsanalyse. In: Möhring, W., & Schlütz, D. (Eds.), *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (pp. 289–306). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-531-18776-1_16

Schneider, G. (2014). Automated media content analysis from the perspective of computational linguistics. In K. Sommer, J. Matthes, M. Wettstein, & W. Wirth (Eds.), *Automatisierung in der Inhaltsanalyse* (pp. 40–54). Köln: von Halem. https://doi.org/10.5167/uzh-108375

Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. *Nature*, 614(7947), 214–216. https://doi.org/10.1038/d41586-023-00340-6

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (1962). The general inquirer: A computer system for content analysis and retrieval based on the sentence as a unit of information. Behavioral Science, 7, 484–498 https://doi.org/10.1002/bs.3830070412

Strippel, C., Bock, A., Katzenbach, C., Mahrt, M., Merten, L., Nuernbergk, C., Pentzold, C., Puschmann, C., & Waldherr, A. (2018). Die Zukunft der Kommunikationswissenschaft ist schon da, sie ist nur ungleich verteilt. Eine Kollektivreplik auf Beiträge im ‘Forum’ (Publizistik Heft 3 und 4, 2016). *Publizistik*, 63, 11–27. https://doi.org/10.1007/s11616-017-0398-5

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An Examination of the Use of Large Language Models to Aid Analysis of Textual Data. *International Journal of Qualitative Methods*, 23. https://doi.org/10.1177/16094069241231168

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

Törnberg, P. (2023b). How to use Large Language Models for Text Analysis. arXiv: 2307.13106. https://doi.org/10.48550/arXiv.2307.13106

Trumbo, C. W. (2004). Research methods in mass communication research: A census of eight journals 1990–2000. Journalism & mass communication quarterly, 81(2), 417-436. https://doi.org/10.1177/107769900408100212

Wiesner, D. (2024, September 25). Politicized or neglected? The role of scientific knowledge in parliamentary debates [Paper presentation]. 10th European Communication Conference (ECREA), Ljubljana, Slovenia. https://flore.unifi.it/retrieve/cb6590cb-c6f1-4442-8592-c006efbc3c8b/ECREA-2024-Abstract-Book.pdf

Wirth, W., Sommer, K., Wettstein, M., & Matthes. J. (2015). Qualitätskriterien in der Inhaltsanalyse. Ein Vorwort. In Wirth, W., Sommer, K., Wettstein, M., & Matthes. J. (Eds.), *Qualitätskriterien in der Inhaltsanalyse* (pp. 9-14). Köln: Halem. https://download.e-bookshelf.de/download/0003/8568/39/L-G-0003856839-0013613709.pdf

Xiao, Z., Yuan, X., Liao, Q. V., Abdelghani, R., & Oudeyer, P. Y. (2023). Supporting qualitative analysis with Large Language Models: Combining codebook with GPT-3 for deductive coding. *arXiv preprint* arXiv:2304.10548. https://doi.org/10.48550/arXiv.2304.10548

Zambrano, A.F., Liu, X., Barany, A., Baker, R.S., Kim, J., Nasiar, N. (2023). From nCoder to ChatGPT: From Automated Coding to Refining Human

Coding. In: Arastoopour Irgens, G., Knight, S. (Eds), *Advances in Quantitative Ethnography. ICQE 2023. Communications in Computer and Information Science*, vol 1895. Springer, Cham. https://doi.org/10.1007/978-3-031-47014-1_32

Acknowledgements

We extend our sincere gratitude to the team of student coders whose diligent work in actor classification provided the essential baseline for our study. Their commitment and expertise were crucial in enabling us to accurately evaluate the performance of ChatGPT.

The authors would also like to mention that they used ChatGPT (version o1-preview) to assist with the preparation of this manuscript. As non-native English speakers, they utilized ChatGPT to translate phrases from German to English and to compose initial drafts for selected sections based on prewritten text segments and notes. Additionally, ChatGPT aided in refining the text to enhance precision, clarity, and coherence, and in correcting grammar and spelling errors. The tool also served as a valuable sparring partner by posing targeted questions that helped to identify missing information crucial for understanding the research presented in this paper. After each use of this tool, the authors thoroughly reviewed and edited the content as needed. The authors take full responsibility for the content of this publication.

Appendix

In this section, we provide background information on the search strings used for selecting the media sample and present the prompts used within the scope of this study. Additionally, we make the raw data and software code available for other researchers. These details offer insights into our research process, enhancing transparency and facilitating the reproducibility of our analyses.

Search Strings

We obtained the media sample for our quantitative content analysis task from the Nexis Uni database, using specific search strings tailored to each of the four science-related issues, namely biotechnology, climate change, neuroscience, and antibiotic resistance (see Table 6).

Issue	Search string
Biotechnology	((Biotech) AND ((gentech) OR (genmani) OR (genet) OR (genom) OR (synthetisch W/1 Biologie) OR (DNA) OR (RNA) OR (Zellkultur ×) OR (biosens) OR (biokataly) OR (gentrans) OR (stammzell) OR (molekular) OR (Mutation) OR (mutier) OR (klon) OR (biomed ×) OR (Genschere) OR (Crispr) OR (Gentherapie) OR (Zellkern) OR (embryo) OR (in-vitro) OR (ips) OR (Keimbahn) OR (transgen) OR (biorak) OR (Stoffwechsel) OR (enzym) OR (ferment) OR (bakteri) OR (protein) OR (mikrob) OR (glucose) OR (molekül) OR (molekuel) OR (kataly) OR (Biokraftstoff) OR (Biotreibstoff) OR (in W/1 vitro) OR (amino) OR (biosicherheit) OR (GMO) OR (GVO) OR (DNS ×) OR (Zelltherapie) OR (biologisch W/1 Sicherheit) OR (rot W/1 Gentechnik) OR (weiß W/1 Gentechnik) OR (grün W/1 Gentechnik))
Climate change	(((klima) AND NOT (Klimaanlage) AND NOT (Klimatisier) AND NOT (Klimax) AND NOT (Klimatechnik) AND NOT (Betriebsklima) AND NOT (Unternehmensklima)) AND ((Klimawandel ×) OR (Klimakrise) OR ((Treibhaus) OR (Erderwärmung) OR (Erderwaermung) OR (globale W/1 Erwärmung) OR (globale W/1 Erwaermung) OR (Klimaziel) OR (CO2) OR (Kohlendioxid) OR (Kohlenstoffdioxid) OR (Luftverschmutz) OR (Umweltverschmutz ×) OR (temperatur) OR (Methan) OR (Klimakatastrophe) OR (Klimatrend) OR (Klimaveränderung) OR (Klimaveraenderung) OR (Klimaänd) OR (Klimaaend) OR (Klimaforsch) OR (Strahlungsantrieb) OR (Klimazustand) OR (Klimawechsel) OR (Klimasystem) OR (Klimaschwank) OR (Weltklima) OR (Extremwetter) OR (Hitzetage) OR (Klimaschutz) OR (Erderhitz) OR (Klimareport) OR (Klimabilanz) OR (klimabericht) OR (klimaneutral) OR (Klimaminist ×) OR (Klimagipfel) OR

	(klimaschädlich) OR (klimaschaedlich ×) OR (Klimaschütz) OR (Klimaschuetz) OR (Dürre) OR (Duerre) OR (Ökosystem) OR (Oekosystem) OR (Biomasse) OR (umweltpoliti ×) OR (klimapoliti) OR (emission) OR (Stickstoff) OR (Meereis) OR (Meeresspiegel) OR (Klimagas) OR (Klimamodell) OR (Klimapaket) OR (Wetterextrem) OR (Klimaverschiebung)))
Neuroscience	(((neuro) AND NOT (neurotisch) AND NOT (pflege) AND NOT (gesundheit)) AND ((Hirn) OR (kognit) OR (Magnetresonanz ×) OR (Kernspint) OR (schizophren) OR (zerebral) OR (cortex) OR (biomarker) OR (tomogra) OR (EEG) OR (pschopatho) OR (bildgebend W/1 Verfahren) OR (magnetstimul) OR (elektrostimul) OR (cognitive) OR (Gehirndop) OR (Alzheimer) OR (Parkinson) OR (Multiple W/1 Sklerose) OR (ADHS) OR (physiolog) OR (stoffwechsel) OR (gedächtnis) OR (gedaechtnis) OR (protein) OR (elektrophysio ×) OR (zyto) OR (zell) OR (Reaktionszeit) OR (Zentralnerven) OR (erinnerung) OR (bewusstsein) OR (Computerchip) OR (implant) OR (wahrnehmung) OR (elektrische W/2 stimul)))
antibiotic resistance	antibio AND resist

Table 6: Search strings for the four science-related issues.

Prompts

To assess the coding performance of ChatGPT, different types of prompts were tested: The first prompt consists of the codebook created for manual content analysis (see Table 7). Furthermore, an optimized prompt version was tested in a stand-alone setting as well as within the NERC pipeline, relying on few-shot learning principles by supplying the model with category keywords instead of providing exhaustive definitions.

Prompting strategy	**Prompt**
Codebook	**Pretext:** Ich werde dir Zeitungsartikel senden. Deine Aufgabe ist es, in den Texten alle namentlich erwähnten Akteure zu finden und diese einem Gesellschaftsbereich zuzuordnen. Sende mir als Output ein valides JSON Array aus Objekten im folgenden Format: [{"gesellschaftsbereich“: "Wissenschaft, Politik, wissenschaftliche Administration, Medizin, Interessensverbände oder

Sonstiges", "name“: "Vorname und Nachname des Akteurs", "nationalität": "Land, in dem die Person arbeitet", "geschlecht": "männlich oder weiblich" }]
Wenn kein Akteur gefunden wird oder der Name nicht bekannt ist, sende bitte ein leeres Array: []

Coding instruction:

Der Gesellschaftsbereich wird an der Institution, für die den Akteur arbeitet, festgemacht.

Die Kategorie Wissenschaft umgreift Forscher:innen ohne politische oder soziale Funktionen. „Wissenschaftler“, „Forscher“ oder „Biologe“ sind eindeutig wissenschaftliche Akteure. Akteure der DFG sind ebenso eindeutig wissenschaftlicher Akteure. Auch Mitglieder der IPCC zählen zu den Wissenschaftlern. Ein Definitionskriterium für wissenschaftliche Akteure ist, dass diese unabhängig/ objektiv arbeiten, also nicht im engeren Sinne interessengeleitet.
ACHTUNG: Wenn bei einem Mediziner erkennbar wird, dass er in seiner Rolle als Wissenschaftler spricht (also forscht und nicht praktiziert), ist die Person als Wissenschaftler zu codieren. Wenn es nicht klar erkennbar ist, dann wird sie nicht als Wissenschaftler codiert. Mitarbeiter von Universitätskliniken (ausgenommen Pflegepersonal, Verwaltungspersonal), wie Chef- oder Oberärzte, werden als wissenschaftliche Akteure
codiert.
ACHTUNG II: Auch Mitarbeiter privater Forschungsinstitute werden den wissenschaftlichen Akteuren zugerechnet – NICHT aber Mitarbeiter von wirtschaftlich orientierten Unternehmen, die auch Forschung betreiben.
ACHTUNG III: Mitarbeiter von Botanischen Gärten werden auch unter Wissenschaft codiert.

Zum politischen Bereich gehören explizit politische Akteure wie Mitglieder von Regierungsinstitutionen, politischer Administration (z.B. Ministerien) sowie politischer Parteien. „CDU Mitglieder“ oder „EU-Abgeordnete“ sind politische Akteure. Auch ehemalige Politiker werden unter Politik kodiert.

Wissenschaftliche Administration beschreibt die etwas engere Klasse von Mitgliedern wissenschaftlichen Institutionen, die auch administrative Funktionen ausführen. Dazu zählen die oben schon erwähnten Ressortforschungseinrichtungen, die einem Bundes- oder Landesministerium unterstellt sind, z.B. – um die Wichtigsten zu

nennen – das Robert Koch Institut, das Bundesamt für Risikobewertung, das Friedrich Löffler Institut, das Paul Ehrlich Institut oder das Bundesamt für gesundheitlichen Verbraucherschutz. Dazu gehören auch Mitglieder internationaler Institutionen wie der WHO oder der ECDC oder des amerikanischen CDC (Centers for Disease Control) oder des NIH (National Instituts of Health). Medizin bezieht sich auf medizinische Fachleute, nämlich Ärzte, nicht auf anderes Krankenhauspersonal allgemein (dies würde unter Sonstiger Bereich codiert werden).
ACHTUNG: Wenn eine Person in einem Artikel nur als Mediziner bezeichnet wird, codieren wir diese als Medizin (nicht als Wissenschaft).

Interessensverbände: Wir unterscheiden zwischen Interessenverbänden, die Kollektivgüter vertreten, wie etwa Umweltschutz, Tierschutz, Frieden, von solchen Interessenverbänden, die die Interessen bestimmter gesellschaftlicher Gruppen vertreten. Zu den Interessenverbänden gehören etwa Greenpeace, Nabu, WWF, auch NGOs. Auch Akteure, die so genannte Kollektivgüterinteressen vertreten, also etwa Umweltschutz etc. zählen zu Interessensverbänden, genau wie Mitglieder von Gewerkschaften, Kirchen, Vertreter von Wirtschaftsunternehmen einschließlich der Pharmaunternehmen und dergleichen, ebenso von Patientenorganisationen. Es wird nicht unterschieden, ob es sich um nationale oder internationale Akteur:innen handelt.
ACHTUNG: Auch Mitarbeiter privater Unternehmen sind bei den Partialinteressenvertreten zu verorten.

Sonstiges: Wenn der Bereich des Akteurs nicht erkennbar ist (wenn z.B. einfach „Expert:innen“ zitiert werden), sind diese Akteue als Sonstige zu codieren. Auch, wenn keiner der zuvor genannten Bereiche zutreffend erscheint, wird Sonstiges codiert. Beispiele sind etwa „Museen“ oder „Zoos“.

Nationalität: Hier wird erfasst, ob es sich bei dem Akteur um eine Person, die (überwiegend) in Deutschland tätigt ist/ arbeitet handelt, oder um eine Person, die in einem anderen Land als Deutschland tätig ist.
ACHTUNG I: Es geht bei dieser Variable *nicht* um die Nationalität (Staatsbürgerschaft) der Person, sondern um ihren aktuellen Tätigkeitsort, bei Wissenschaftlern etwa, ob sie an einer deutschen Hochschule forschen oder nicht. Oder handelt es sich um einen Politiker aus dem deutschen Bundestag, oder nicht.

Optimized prompt and NERC pipeline	**Pretext for the optimized prompt (stand-alone version):** Du bist ein Experte für die Extraktion von Informationen aus Texten. Du extrahierst Eigenschaften von natürlichen Personen. Sende mir ein valides JSON Array im folgenden Format: ["name": "Vorname und Nachname", "geschlecht": "männlich oder weiblich", "gesellschaftsbereich": "Wissenschaft, Politik, wissenschaftliche Administration, Medizin, Interessensvertretung oder Sonstiges", "tätigkeit": "genannter Beruf und Unternehmen / Verband", "nationalität": "Land, in dem die Person arbeitet"] Wenn keine Person gefunden wird oder der Name nicht bekannt ist, sende bitte ein leeres Array: []
	Pretext for the NERC pipeline: Du bist ein Experte für die Extraktion von Informationen aus Texten. Suche Informationen zur Person [VORNAME NACHNAME].
	Coding instruction: Gesellschaftlicher Bereich: Orientiere dich an der Tätigkeit und Institution, an der die Person arbeitet. 1. Wissenschaft: Z.B. Wissenschaftler, Forscher, Biologen, Doktoranden, Akteure der DFG (Deutsche Forschungsgesellschaft) oder des IPCC (Intergovernmental Panel on Climate Change), Angestellte an Botanischen Gärten, forschende Mediziner, Chefärzte und Oberärzte an Universitätskliniken. 2. Politik: Z.B. Mitarbeiter von Ministerien, Gesundheitsämtern, UNESCO oder FAO (Food and Agriculture Organization) und Regierungsvertreter, Staatssekretäre, Diplomaten, Parteiangehörige, Parteichefs, Parlamentarier sowie Mitglieder der Europäischen Union (EU), der Bundesregierung, Behörden der Bundesländer, Opposition, von Fraktionen, Bürgermeister, Parteimitglieder von CDU, CSU, SPD, Grüne, FDP, AfD, Linke/PDS. 3. Wissenschaftliche Administration: Z.B. Mitarbeiter der Ressortforschung an Bundesministerien und Landesministerien 4. Medizin: Z.B. Praktizierende Ärzte, Mediziner, Fachärzte, Neurologen. 5. Interessensvertretung: Z.B. Mitarbeiter von Verbänden, Stiftungen, Greenpeace, Nabu, WWF, NGOs, Gewerkschaften, Kirchen, Firmen, Patientenorganisationen, Privatwirtschaft und Lobbygruppen. Gründer, Manager, Vorstandsmitglieder, Geschäftsführer, Unternehmer. Wissenschaftler bei Pharmakonzernen und anderen

	Firmen, Industrieforscher und Forschungsdienstleister. 6. Sonstiges: Personen, zu die zu keiner der vorherigen Kategorien gut passen oder zu denen ein anderes Label besser passt. Z.B. Pflegepersonal, Patienten, Museen, Zoos, Journalisten, Autoren, Lehrer, Film und Fernsehen, Kulturschaffende. Ebenfalls Personen, für die keine spezifischen Informationen zur beruflichen Tätigkeit oder Zugehörigkeit zu einem bestimmten Bereich vorliegen.

Table 7: Prompts for the tested prompting strategies.

Data and Code

In adherence to the principles of Open Science and transparency, we make both the code used to prompt ChatGPT and the datasets, coded manually and through automated content analysis with ChatGPT, accessible through a Git repository: https://gitlab.com/wisskomm-in-digitalen-medien/chatgpt_in_quantitative_content_analysis_nerc