A JOURNAL BY SISSA MEDIALAB

JCom

SPECIAL ISSUE
Science Communication in the Age of Artificial Intelligence

ARTICLE

ChatGPT's potential for quantitative content analysis: categorizing actors in German news articles

Clarissa Elisabeth Hohenwalde, Melanie Leidecker-Sandmann, Nikolai Promies and Markus Lehmkuhl

Abstract

We assessed ChatGPT's ability to identify and categorize actors in German news media articles into societal groups. Through three experiments, we evaluated various models and prompting strategies. In experiment 1, we found that providing ChatGPT with codebooks designed for manual content analysis was insufficient. However, combining Named Entity Recognition with an optimized prompt for actor Classification (NERC pipeline) yielded acceptable results. In experiment 2, we compared the performance of gpt-3.5-turbo, gpt-4o, and gpt-4-turbo, with the latter performing best, though challenges remained in classifying nuanced actor categories. In experiment 3, we demonstrated that repeating the classification with the same model produced highly reliable results, even across different release versions.

Keywords

AI tools in science communication; Science and media

Received: 30th October 2024
Accepted: 27th February 2025
Published: 14th April 2025

1 ▪ 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; 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 fragmented digital content in today's 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]. Compared to manual coding, automated coding relies primarily on computational resources, making big data analyses more manageable, facilitating real-time analysis, and offering significant financial savings [e.g. Scharkow, 2013]. However, 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, 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 (NLP) capabilities make them particularly interesting for content analysis tasks.

A typical task in content analysis is the identification and classification of actor groups within journalistic news 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. 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 communication and social 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.

2 ▪ Advancements of automated content analysis in communication science

Automated content analysis has a long-standing tradition in communication science, predating the development of LLMs. Already in the early 1960s, dictionary approaches were introduced [Stone et al., 2007] that enabled 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, 2015], 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 (BoW) models and have been shown to outperform dictionaries in various contexts [Burscher et al., 2015; Kroon et al., 2024; Scharkow, 2011]. In supervised approaches, machine learning models are trained on manually labeled data to inductively develop statistical models that aim to replicate human coding decisions [Brosius et al., 2022; Scharkow, 2011]. This method allows for a more nuanced handling of textual data characteristics and helps mitigate researcher bias by enabling the algorithm to detect patterns that guide classification [Burscher et al., 2015; Chew et al., 2023; Kroon et al., 2024]. However, supervised machine learning relies on large volumes of manually annotated training data to ensure validity. This requirement is particularly challenging in social science research, where new studies often demand domain-specific training datasets tailored to their specific research questions [Chew et al., 2023; Laurer et al., 2024; Törnberg, 2024a]. Moreover, these supervised methods do not generalize well across different languages, domains, or genres, further limiting their utility [Kroon et al., 2024]. To address these challenges, pretrained models have been developed, offering a foundational linguistic understanding that can be fine-tuned for specific tasks [Brosius et al., 2022]. However, adapting or fine-tuning pretrained models still requires substantial programming expertise and access to large, labeled datasets.

The emergence of LLMs offers new possibilities: Transformer-based models are trained on large volumes of unstructured text data, allowing them to acquire transferable language knowledge that can be applied to various downstream tasks with minimal task-specific training data [Brown et al., 2020; Kroon et al., 2024; Törnberg, 2024a]. 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 previously unseen categories by interpreting class descriptors provided in the prompt [Brown et al., 2020]. Few-shot learning enhances 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.

3 · Applications of LLMs in social sciences — state of research

Various scholars have proposed potential applications for LLMs in social science research [Argyle et al., 2023; Bail, 2024; Binz & Schulz, 2023; Stokel-Walker & Van Noorden, 2023], and automated methods are becoming increasingly prominent in this field. As a result, the need for methodological discussions about the quality requirements, validity and reliability of individual methods is growing [Buz et al., 2022; Niekler, 2018], a gap that our study seeks to address. It is crucial to assess how well LLMs like ChatGPT perform in comparison to

established human-driven methods, how they navigate task-specific applications in content analysis, what factors influence their performance, and to consider the broader implications of their use in social science research.

3.1 ▪ *Comparison of ChatGPT's performance with human coders*

Multiple studies have evaluated ChatGPT's ability to perform coding tasks in comparison to human annotators, revealing both its strengths and limitations.

Alomari [2024] conducted a systematic review of ChatGPT's performance across various NLP tasks, including Named Entity Recognition (NER) and text classification. The study concludes that ChatGPT demonstrates significant potential in enhancing NLP workflows, as ChatGPT effectively comprehends diverse contexts and extracts meaningful details.

With regards to topic, frame and tone classifications, among others, research by Gilardi et al. [2023] showed that ChatGPT-3.5-turbo achieved higher accuracy for most topic and frame classification tasks in news articles and tweets than Amazon Mechanical Turk (MTurk) crowdworkers compared to a baseline of trained coders. Li et al. [2024] compared ChatGPT-3.5-turbo's performance in detecting hateful, offensive, and toxic comments on social media against human-coded annotations and found an 80 % accuracy rate, noting that ChatGPT was even able to correctly identify tweets written in an implicit fashion.

In the domain of political text analysis, Heseltine and Clemm von Hohenberg [2024] demonstrated that ChatGPT-4's coding was highly accurate for the variables "political content", "negativity", "sentiment" and "ideology" when compared to human expert coders. They further suggested a hybrid approach in which disagreements between multiple GPT-4 runs were resolved by human experts to enhance accuracy. Törnberg [2024b] found that ChatGPT-4 was capable of accurately inferring political affiliations from social media content, even outpacing human expert coders in some cases.

3.2 ▪ *Task-specific applications of ChatGPT in content analysis*

Beyond direct comparisons to human coders, several studies have examined ChatGPT's application in specific content analysis tasks.

Regarding the automatic classification of actors in texts, 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 study successfully applied a dictionary-based segmentation strategy followed by a ChatGPT-based analysis (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.

Leas et al. [2024] successfully identified adverse events[1] in social media posts using ChatGPT, suggesting its potential to replace human coders. Gielens et al. [2025] tested ChatGPT-4-turbo and ChatGPT-4 in classifying predefined arguments in policy debates and found high accuracy and reliability, though they recommended further validation before widespread implementation. Zambrano et al. [2023] assessed ChatGPT-4's ability to classify socially positive and negative constructs in press-related texts, reporting strong performance in clear categorization tasks but a tendency to overgeneralize in more ambiguous cases. Huang et al. [2023] found that ChatGPT accurately identified implicit hateful tweets. Additionally, they observed that in cases of disagreement, ChatGPT's classifications aligned more closely with laypeople's perceptions.

3.3 ▪ *Factors influencing ChatGPT's performance*

Research has also already identified several key factors that influence ChatGPT's effectiveness in coding tasks:

Prompt Design: Xiao et al. [2023] and Tai et al. [2024] demonstrated that well-structured prompts with clear codebooks and examples yielded results comparable to manual coding. Providing contextual information in the form of "code — description — examples" significantly improved reliability. Aldeen et al. [2023] further highlighted that the effectiveness of prompt strategies can vary depending on the nature of the dataset and the specific labeling task. Kim and Lu [2024] showed that using the prompting technique of few-shot learning and applying prompt refinement led to small improvements in ChatGPT's performance in rhetorical move-step analysis, while fine-tuning led to substantial gains in accuracy. Alizadeh et al. [2025] further demonstrated that fine-tuning significantly improves the performance of open-source LLMs, often allowing them to match or surpass zero-shot GPT-3.5 and GPT-4. Contrary, Kuzman and Ljubešić [2023] found that in a zero-shot setting, ChatGPT-3.5 performed better than a fine-tuned transformer-based model in genre classification tasks. Beyond manually crafted prompts, automatic prompt optimization has been proposed as a method to systematically generate high-quality prompts [Abraham et al., 2024].

Text Characteristics and Language: Heseltine and Clemm von Hohenberg [2024] found that performance is higher for short texts like tweets compared to longer articles and that Non-English text presents additional challenges.

Hybrid Approaches — Verification by Human Coders: Heseltine and Clemm von Hohenberg [2024] advocated for hybrid approaches, where human coders resolve discrepancies between multiple iterations[2] of coding to enhance accuracy. Yu [2025] demonstrated that integrating LLMs into the annotation of CEO statements can significantly reduce researchers' workload, though human verification remains necessary to ensure reliability.

Considering this state of research, we conclude that while ChatGPT shows considerable potential for some coding tasks, its effectiveness is inconsistent across different

1. In medical research, an adverse event refers to any unintended or harmful occurrence that happens during or after a medical treatment, clinical trial, or intervention, regardless of whether it is directly caused by the treatment. Such events can include physical symptoms, abnormal laboratory findings, or psychological effects.
2. Tai et al. [2024] found that performing multiple iterations of coding with ChatGPT improved result consistency.

applications. Although LLMs often match human performance in text analysis, their reliability differs depending on material, language, and prompt structure [Ollion et al., 2023]. Moreover, different LLMs vary in their coding performance. Ziems et al. [2024], for example, evaluated the zero-shot performance of various language models on 24 social science coding tasks, finding considerable variation in accuracy — ranging from human-comparable levels to near-random baselines. These limitations highlight the need for further research to refine LLM-based coding approaches and ensure their robustness across different analytical contexts.

4 ▪ Challenges in actor recognition and classification

The analytical context of our study is the identification and categorization of societal actors in journalistic news media articles, which is a crucial task in communication research. Analyzing the structure of actors in public discourse provides valuable insights into which voices are represented in public debates and how different societal groups participate in shaping public opinion and policy-making [Leidecker-Sandmann & Lehmkuhl, 2022].

From a methodological perspective, manually coding actors in journalistic texts is a labor-intensive and complex task. Many empirical studies examine the diversity of actors in news media coverage by analyzing how frequently specific actors appear and under what circumstances and thereby contribute to our understanding of news media influence on public debate [e.g. Leidecker-Sandmann & Lehmkuhl, 2022; Leidecker-Sandmann et al., 2022; Burggraaff & Trilling, 2017; Eisenegger et al., 2020; Maurer et al., 2021; Niekler, 2018]. However, the ambiguity and contextual nature of actor references pose an ongoing challenge (e.g. that formulations such as “he/she says” do not automatically make it clear which person is meant by “he/she”).

Over time, automated approaches such as NER have been developed in computational text analysis. NER techniques extract named entities, such as persons, organizations, and location, from unstructured text and have proven effective in journalistic contexts [Buz et al., 2022; Marrero et al., 2013; Schneider, 2014]. However, the quality of NER output varies depending on factors such as language, textual genre, and entity type [Nadeau & Sekine, 2009]. Standard NER methods are also inherently limited to identifying entities without assigning them to broader societal categories. The combined task of Named Entity Recognition and Classification (NERC), an important subtask of Information Extraction (IE), attempts to bridge this gap by assigning identified entities to relevant societal groups [Goyal et al., 2018].

5 ▪ 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, and reliability.

5.1 ▪ *Sample description*

As a case study, we chose a typical task of quantitative content analysis, namely to identify and categorize actor groups into different societal groups in science-related public 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 covered both mainstream and regional legacy media outlets within the German news 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 news 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 amounts of news 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).

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

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
		∑ **2,883**

5.2 ▪ *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 NER approach. The NER tool detected which words in the articles constituted a person's name. We performed this process using the Python-based package FLAIR [Akbik et al., 2019], which has demonstrated high precision and recall in previous validations for German journalistic texts. In a previous analysis [Buz et al., 2022], it accurately identified 99 % of relevant individual names with minimal errors, forming a strong foundation for the subsequent classification of these actors into societal groups.

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 the following actors in the public decision-making process 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 $\alpha = .63$, indicating a rather low level of agreement. 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 $\alpha = .77$, which is considered acceptable for tentative conclusions in content analysis.

The moderate agreement levels can be explained by 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, highlighting the challenge of accurately categorizing actors based solely on textual mentions.

Following the reliability assessment and coding scheme adjustment, the coding resulted 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.

5.3 ▪ *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 (November 2023), gpt-3.5-turbo was the latest model available from OpenAI. We explored three approaches to prompt the model:[3]

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 must 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. This means that instead of providing exhaustive definitions, we supplied the model with category keywords as 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 NERC pipeline to reduce task complexity. We performed automatic 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.

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, ensuring predictable and controlled results by consistently selecting the highest-probability response and eliminating randomness. We compared ChatGPT's classifications against our human-coded dataset, which served as the gold standard.[4] Due to the multi-class classification problem with an imbalanced class distribution, we assessed overall performance using the macro-averaged F1-score. This approach balances precision and recall across all classes, treating each class equally regardless of its size. To further substantiate our findings in cases where the results provided valuable insights to strengthen

3. The full prompts can be found in the appendix.
4. 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).

the overall argument, we analyzed class-specific F1-scores, precision, recall, and the confusion matrix. This allowed us to identify categories with frequent misclassifications and to pinpoint areas where category distinctions required further refinement.

5.4 ▪ *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 to ensure consistency across models. We conducted API requests for gpt-4-turbo in November 2023 and for gpt-4o in September 2024. The performance of each model was again evaluated against the human-coded gold standard using the macro-averaged F1-score. 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.

5.5 ▪ *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. As far as we know, most studies usually only focus on the changes between different models, but do not analyze changes between different releases of one and the same model. However, we believe, 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. 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. This led to 20 classifications per identified actor. All API requests were conducted in September 2024.

To determine reliability within a single release, we calculated Krippendorff's alpha based on the agreement between the ten classifications produced by this release. To determine reliability across releases, we calculated Krippendorff's alpha using all 20 classifications per actor, treating the classifications from both versions as separate coders.

In this case, we also calculated confidence intervals to estimate the variance of the reliability coefficients. To do this, we created 1,000 bootstrap samples for each reliability test by resampling with replacement from the set of classified actors and then calculated Krippendorff's alpha for each sample. As a result, we obtained 1,000 Krippendorff's alpha values for each of the three comparisons (for the first and second intra-release reliability and the inter-release reliability). The confidence intervals were determined from the 2.5 and 97.5 percentiles of these values. This approach provided robust estimates of reliability and allowed us to assess the stability of classifications over time.

5.6 ▪ *Statistical analysis*

For the statistical analysis of the 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.

6 ▪ Results

6.1 ▪ *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 NERC pipeline.

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

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**

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 with accurately identifying actors in the "Advocacy" and "Other Actors" categories. This resulted in low precision and recall for these categories.

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

		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

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 Thomas Schulze is classified as "Science", despite working for a company and therefore representing partial interests:[5]

> "'With our method, information about the origin of an animal, i.e. country, farm, herd and even BSE test result, can be securely encrypted,' explains Thomas Schulze, head of the research and development department at November AG.".

Table 4. Confusion matrix for the integration of ChatGPT into the NERC pipeline. The number of actors differs from Table 3, as more actors were identified by NER than in the setting utilizing the codebook.

		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

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

5. Translation by the authors. Original text: "'Mit unserer Methode können Informationen über die Herkunft eines Tieres, also Land, Hof, Herde und gar BSE-Testergebnis, sicher verschlüsselt werden', erklärt Thomas Schulze, Leiter der Forschungs- und Entwicklungsabteilung der November AG."

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".[6] Here, the CEO Detlev Goj was classified as "Other Actors", even though he represents his company and should therefore be classified as "Advocacy".

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

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.

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

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

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"

6. Translation by the authors. Original text: "Durchschnittlich 400 bis 500 Pakete mit sterilen Maden versendet Firmenchef Detlev Goj monatlich."

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”.

6.3 ▪ *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, 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.

7 ▪ Discussion

This study examined ChatGPT’s potential to replace human coders in identifying and categorizing actor groups within German news media coverage as part of a quantitative content analysis. Through three experiments, we assessed the validity, performance, and reliability of different ChatGPT models and prompting strategies.

Experiment 1 demonstrated that using the original codebook designed for human coders was insufficient for achieving valid actor recognition and classification. Especially the categories “Advocacy” and “Other Actors” posed significant challenges. We found that advocacy actors were frequently misclassified as “Science”. 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 NERC pipeline substantially enhanced performance. 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 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. 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. This finding underscores the continued relevance of our study, even as newer models are released.

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 and August 6 release of gpt-4o, with overlapping 95 % confidence intervals. 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 model releases help mitigate concerns about the impact of updates on research reproducibility.

7.1 ▪ *Limitations*

Several limitations of this study should be acknowledged. First, our analysis focused exclusively on German-language news media articles and four science-related public debates. This scope may limit the generalizability of our findings to other languages, topics, or text types. Second, the actor classification scheme was simplified due to low inter-coder reliability among human coders. While ChatGPT's coding was benchmarked against manual coding treated as the gold standard, it is important to recognize that this standard itself is not without errors. The observed intercoder reliability of Krippendorff's alpha $\alpha = .77$ indicates moderate agreement, highlighting the inherent challenges of actor classification, both for human coders and automated systems. Third, a potential limitation of our classification scheme concerns the heterogeneous nature of advocacy groups. As defined in this study, the category includes a broad spectrum of actors, ranging from NGOs and trade unions to representatives of commercial enterprises, patient organizations, and even lobbyists. While this classification allows for a comprehensive mapping of actors engaging in public discourse, it also introduces a certain degree of conceptual vagueness. Fourth, the study faced challenges related to imbalanced class distribution, which affected both human and automated classification. To mitigate this issue, we calculated macro-averaged F1-scores; however, the low number of actors in certain categories may still have biased the performance metrics. A fifth limitation concerns model selection. We tested only one company's models — ChatGPT by OpenAI — which, while 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 could offer better performance or different classification capabilities, warranting further comparative evaluation. Additionally, OpenAI's models are proprietary, meaning that access, availability, and cost are controlled by the company. This raises concerns about the replicability of studies conducted using ChatGPT. Finally, relying on pretrained models without domain-specific fine-tuning may have constrained the model's ability to capture the specialized knowledge needed for accurate classification.

8 ▪ 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 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:

- Structure the Prompt: Clearly define the LLM's role and task to ensure more precise and contextually appropriate responses.
- Use Structured Outputs: Instead of describing the desired format within the prompt, enforce adherence to a predefined JSON schema. This approach improves type safety, facilitates automatic parsing, and ensures consistent, machine-readable outputs.
- 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. 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]. This challenges the assumption of linear model improvement and highlights the need for empirical evaluation over reliance on version numbers. We therefore recommend that model selection should be guided by task-specific evaluations rather than assumptions based on model updates.

Notably, ChatGPT's high reliability underscores its suitability for longitudinal research, where consistency is essential. 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 the 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 NLP tools to further enhance performance.

We believe that ChatGPT has significant potential as a valuable tool for researchers in journalism studies and related fields, providing notable 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.

Acknowledgments

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. Furthermore, we thank the anonymous reviewers for their insightful suggestions that helped us improve this paper.

We would also like to mention that we used ChatGPT (version o1-preview) to assist with the preparation of this manuscript. As non-native English speakers, we 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, we thoroughly reviewed and edited the content as needed. We take full responsibility for the content of this publication.

A ▪ Search strings, prompts, data and code

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.

A.1 ▪ *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).

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

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

A.2 ▪ *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 for the original German version and Table 8 for an English translation). 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.

Table 7. Original prompts for the tested prompting strategies (German).

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: []

Continued on the next page.

Continued from the previous page.

Prompting strategy	Prompt
	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.

Continued on the next page.

Continued from the previous page.

Prompting strategy	Prompt
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 8. English translation of the prompts for the tested prompting strategies.

Prompting strategy	Prompt
Codebook	**Pretext:** I will send you newspaper articles. Your task is to find all people mentioned by name in the texts and assign them to a social sector. Send me as output a valid JSON in the following format: [{"social sector": "science, politics, scientific administration, medicine, advocacy groups or other", "name": "first name and surname of the person", "nationality": "country in which the person works", "gender": "male or female" }] If no person is found or the name is not known, please send an empty array: []

Continued on the next page.

Continued from the previous page.

Prompting strategy	Prompt
	Coding instruction: The social sector is determined by the institution for which the person works. The category "science" encompasses researchers without political or social functions. "Scientist", "researcher" or "biologist" are to be classified as "science". DFG staff and members of the IPCC also are part of "science". One definition criterion for "science" is that scientists work independently/objectively, i.e. that they are not interest-driven in the narrow sense. **NOTE I**: If it is clear that a medical practitioner is speaking in their role as a scientist (i.e. researching and not practicing), the person should be coded as a "science". If it is not clearly recognizable, then the person is not to be coded as "science". Employees of university hospitals (except nursing staff, administrative staff), such as chief physicians or senior physicians, are coded as "science". **NOTE II**: Employees of private research institutes are also counted as "science" — but NOT employees of commercially oriented companies that also conduct research. **NOTE III**: Employees of botanical gardens are also coded under "science". The category "politics" includes political actors such as members of government institutions, political administration (e.g. ministries) and political parties. "CDU member" or "Member of the European Parliament" belong to "politics". Former politicians are also coded under "politics". "Scientific administration" describes the somewhat narrower class of members of scientific institutions that also perform administrative functions. These include departmental research institutions, which are subordinate to a federal or state ministry, e.g. — to name the most important ones — the Robert Koch Institute, the German Institute for Risk Assessment, the Friedrich Löffler Institute, the Paul Ehrlich Institute or the Federal Office of Consumer Protection and Food Safety. This also includes members of international institutions such as the WHO or the ECDC or the American CDC (Centers for Disease Control) or the NIH (National Institutes of Health). The category "medicine" refers to medical professionals, namely doctors, not other hospital staff in general (this would be coded under "other"). **CAUTION**: If a person is only referred to as a medical professional in an article, we code them as "medicine" (not "science"). "Advocacy groups": We distinguish between advocacy groups that represent collective goods, such as environmental protection, animal welfare and peace, and advocacy groups that represent the interests of specific social groups. Advocacy groups include Greenpeace, Nabu, WWF and NGOs. Actors representing so-called collective goods interests, such as environmental protection, etc., are also "advocacy groups", as are members of trade unions, churches, representatives of commercial enterprises including pharmaceutical companies and the like, as well as patient organizations. No distinction is made as to whether these are national or international actors. **ATTENTION**: Employees of private companies are also included as "advocacy groups". "Other": If the societal sector is not recognizable (e.g. if "experts" are simply quoted), these people are to be coded as "other". "Other" is also coded if none of the aforementioned areas appear to apply. Examples include "museums" or "zoos". Nationality: This records whether the person (predominantly) works in Germany or in a country other than Germany. ATTENTION I: This variable is *not* about the nationality (citizenship) of the person, but about their current place of work, for example, in the case of scientists, whether they are doing research at a German university or not. Or is it a politician from the German Bundestag or not.

Continued on the next page.

Continued from the previous page.

Prompting strategy	Prompt
Optimized prompt and NERC pipeline	**Pretext for the optimized prompt (stand-alone version):** You are an expert in extracting information from texts. You extract characteristics of natural persons. Send me a valid JSON array in the following format: ["name": "first name and surname of the person", "gender": "male or female", "field of society": "science, politics, scientific administration, medicine, advocacy groups or other", "activity": "named profession and company / association", "nationality": "country where the person works"] If no person is found or the name is not known, please send an empty array: []
	Pretext for the NERC pipeline: You are an expert in extracting information from texts. Search for information about the person [FIRSTNAME SURNAME].
	Coding instruction: Gesellschaftlicher Bereich: Orientiere dich an der Tätigkeit und Institution, an der die Person arbeitet. Social sector: Focus on the activity and institution where the person works. 1. "Science": e.g. scientists, researchers, biologists, doctoral students, members of the DFG (German Research Foundation) or the IPCC (Intergovernmental Panel on Climate Change), employees at botanical gardens, medical researchers, chief physicians and senior physicians at university hospitals. 2. "Politics": e.g. employees of ministries, health authorities, UNESCO or FAO (Food and Agriculture Organization) and government representatives, state secretaries, diplomats, party members, party leaders, parliamentarians and members of the European Union (EU), the federal government, authorities of the federal states, members of the opposition, parliamentary groups, mayors, party members of the German parties CDU, CSU, SPD, Grüne, FDP, AfD, Linke/PDS. 3. "Scientific administration": e.g. employees of departmental research at federal and state ministries 4. "Medicine": e.g. practicing doctors, physicians, specialists, neurologists. 5. "Advocacy groups": e.g. employees of associations, foundations, Greenpeace, Nabu, WWF, NGOs, trade unions, churches, companies, patient organizations, private industry and lobby groups. Founders, managers, board members, managing directors, entrepreneurs. Scientists at pharmaceutical companies and other companies, industry researchers and research service providers. 6. "Other": People who do not fit well into any of the previous categories or for whom another label fits better. E.g. nursing staff, patients, museums, zoos, journalists, authors, teachers, film and television, cultural workers. Also people for whom there is no specific information on their professional activity or affiliation to a particular social sector.

A.3 ▪ *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.

References

Abraham, L., Arnal, C., & Marie, A. (2024). Prompt selection matters: enhancing text annotations for social sciences with Large Language Models. https://doi.org/10.48550/arXiv.2407.10645

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]. https://github.com/flairNLP/flair

Aldeen, M., Luo, J., Lian, A., Zheng, V., Hong, A., Yetukuri, P., & Cheng, L. (2023). ChatGPT vs. human annotators: a comprehensive analysis of chatGPT for text annotation. *2023 International Conference on Machine Learning and Applications (ICMLA)*, 602–609. https://doi.org/10.1109/ICMLA58977.2023.00089

Alizadeh, M., Kubli, M., Samei, Z., Dehghani, S., Zahedivafa, M., Bermeo, J. D., Korobeynikova, M., & Gilardi, F. (2025). Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, *8*, 17. https://doi.org/10.1007/s42001-024-00345-9

Alomari, E. A. (2024). Unlocking the potential: a comprehensive systematic review of ChatGPT in natural language processing tasks. *Computer Modeling in Engineering & Sciences*, *141*, 43–85. https://doi.org/10.32604/cmes.2024.052256

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*, 337–351. https://doi.org/10.1017/pan.2023.2

Bail, C. A. (2024). Can generative AI improve social science? *Proceedings of the National Academy of Sciences*, *121*, e2314021121. https://doi.org/10.1073/pnas.2314021121

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

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: an overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*, 8–23. https://doi.org/10.1080/21670811.2015.1096598

Brosius, H. B., Haas, A., & Unkel, J. (2022). Inhaltsanalyse III: Automatisierte Inhaltsanalyse. In H.-B. Brosius, A. Haas & J. Unkel (Eds.), *Methoden der empirischen Kommunikationsforschung: Eine Einführung* (pp. 179–194). Springer Fachmedien. 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., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Burggraaff, C., & Trilling, D. (2017). Through a different gate: an automated content analysis of how online news and print news differ. *Journalism*, *21*, 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*, 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. *Studies in Communication and Media*, *10*, 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. 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

Gielens, E., Sowula, J., & Leifeld, P. (2025). Goodbye human annotators? Content analysis of social policy debates using ChatGPT. *Journal of Social Policy*, 1–20. https://doi.org/10.1017/s0047279424000382

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, *120*, e2305016120. 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*, 115–127. https://doi.org/10.14198/medcom000018

Goyal, A., Gupta, V., & Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, *29*, 21–43. https://doi.org/10.1016/j.cosrev.2018.06.001

Günther, E., & Quandt, T. (2015). Word counts and topic models: automated text analysis methods for digital journalism research. *Digital Journalism*, *4*, 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*. Suhrkamp.

Haim, M. (2023). Texte als Daten I. In M. Haim (Ed.), *Computational Communication Science* (pp. 169–193). Springer Fachmedien. https://doi.org/10.1007/978-3-658-40171-9_8

Heseltine, M., & Clemm von Hohenberg, B. (2024). Large language models as a substitute for human experts in annotating political text. *Research & Politics*, *11*, 20531680241236239. https://doi.org/10.1177/20531680241236239

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech. *Companion Proceedings of the ACM Web Conference 2023*, 294–297. https://doi.org/10.1145/3543873.3587368

Kim, M., & Lu, X. (2024). Exploring the potential of using ChatGPT for rhetorical move-step analysis: the impact of prompt refinement, few-shot learning and fine-tuning. *Journal of English for Academic Purposes*, *71*, 101422. https://doi.org/10.1016/j.jeap.2024.101422

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*, 142–162. https://doi.org/10.1080/19312458.2023.2261372

Kuzman, T., & Ljubešić, N. (2023). Automatic genre identification: a survey. *Language Resources and Evaluation*, *59*, 537–570. https://doi.org/10.1007/s10579-023-09695-8

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*, 84–100. https://doi.org/10.1017/pan.2023.20

Leas, E. C., Ayers, J. W., Desai, N., Dredze, M., Hogarth, M., & Smith, D. M. (2024). Using Large Language Models to support content analysis: a case study of ChatGPT for adverse event detection. *Journal of Medical Internet Research*, *26*, e52499. https://doi.org/10.2196/52499

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*, 847–866. https://doi.org/10.1177/09636625221095740

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 and Media*, *11*, 337–393. https://doi.org/10.5771/2192-4007-2022-3-337

Li, L., Fan, L., Atreja, S., & Hemphill, L. (2024). "HOT" ChatGPT: the promise of ChatGPT in detecting and discriminating hateful, offensive and toxic comments on social media. *ACM Transactions on the Web*, *18*, 1–36. https://doi.org/10.1145/3643829

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*, 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-augstein-stiftung.de/wp-content/uploads/2021/11/Studie-einseitig-unkritisch-regierungsnah-reinemann-rudolf-augstein-stiftung.pdf

Nadeau, D., & Sekine, S. (2009). A survey of named entity recognition and classification. In *Named Entities* (pp. 3–28). John Benjamins Publishing Company. https://doi.org/10.1075/bct.19.03nad

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

Niekler, A. (2018). *Automatisierte Verfahren für die Themenanalyse nachrichtenorientierter Textquellen*. 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! 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*. 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*, 515–524. https://doi.org/10.1177/107769909707400306

Saravia, E. (2022). *Prompt engineering guide*. 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. *Quality & Quantity*, *47*, 761–773. https://doi.org/10.1007/s11135-011-9545-7

Scharkow, M. (2013). Automatische Inhaltsanalyse. In *Handbuch standardisierte Erhebungsverfahren in der Kommunikationswissenschaft* (pp. 289–306). Springer Fachmedien. 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). 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*, 214–216. https://doi.org/10.1038/d41586-023-00340-6

Stone, P. J., Bales, R. F., Namenwirth, J. Z., & Ogilvie, D. M. (2007). 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. (2024a). *How to use Large-Language Models for text analysis* [How-to guide]. SAGE Publications Ltd. https://doi.org/10.4135/9781529683707

Törnberg, P. (2024b). Large Language Models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*. https://doi.org/10.1177/08944393241286471

Trumbo, C. W. (2004). Research methods in mass communication research: a census of eight journals 1990–2000. *Journalism & Mass Communication Quarterly*, *81*, 417–436. https://doi.org/10.1177/107769900408100212

Wiesner, D. (2024). Politicized or neglected? The role of scientific knowledge in parliamentary debates [Paper presentation]. *10th European Communication Conference (ECREA)*. https://flore.unifi.it/retrieve/cb6590cb-c6f1-4442-8592-c006efbc3c8b/ECREA-2024-Abstract-Book.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. *28th International Conference on Intelligent User Interfaces*, 75–78. https://doi.org/10.1145/3581754.3584136

Yu, D. (2025). Towards LLM-assisted move annotation: leveraging ChatGPT-4 to analyse the genre structure of CEO statements in corporate social responsibility reports. *English for Specific Purposes*, *78*, 33–49. https://doi.org/10.1016/j.esp.2024.11.003

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 *Advances in Quantitative Ethnography* (pp. 470–485). Springer. https://doi.org/10.1007/978-3-031-47014-1_32

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

About the authors

Clarissa Elisabeth Hohenwalde is a research assistant at the Department of Science Communication at the Karlsruhe Institute of Technology, Germany. Her research focuses on large language models (LLMs) in science communication, computational social science, and research methods for journalism and communication science.

ai3551@kit.edu

Melanie Leidecker-Sandmann (Dr. phil.) is a research assistant at the Department of Science Communication at the Karlsruhe Institute of Technology, Germany. Her research focuses on science communication, political communication as well as on media content and journalism research.

leidecker-sandmann@kit.edu @leidecker-sandmann

Nikolai Promies is a PhD student at the Department of Science Communication at the Karlsruhe Institute of Technology, Germany. His research explores the application of automated methods to analyze patterns in the structure of public discourses, with a focus on science journalism.

nikolai.promies@posteo.de

Markus Lehmkuhl is Professor of Science Communication in Digital Media at the Karlsruhe Institute of Technology, Germany. His research focuses on the emergence and structure of public opinion formation on scientific topics in general and risk topics in particular, with an emphasis on the role of journalism.

markus.lehmkuhl@kit.edu

How to cite

Hohenwalde, C. E., Leidecker-Sandmann, M., Promies, N. and Lehmkuhl, M. (2025). 'ChatGPT's potential for quantitative content analysis: categorizing actors in German news articles'. *JCOM* 24(02), A01. https://doi.org/10.22323/2.24020201.

ISSN 1824-2049. Published by SISSA Medialab. jcom.sissa.it