# Comparing and Modeling Argumentation in German Political Communication across Arenas

**Nina Vikhrova**[1] **and Johannes Kühling**[2] **and Sebastian Haunss**[2] **and Sebastian Padó**[1]
[1]IMS, Universität Stuttgart, [2]SOCIUM, Universität Bremen
{nina.vikhrova|pado}@ims.uni-stuttgart.de, {kuehling|haunss}@uni-bremen.de

## Abstract

Deliberation, involving the formulation and exchange of arguments, forms an integral part of political decision making in democracies. Argumentation patterns however differ substantially across different political arenas, such as plenary speeches and committee meetings. However, despite a lot of interest in argumentation, there is comparatively little computational work on analyzing differences in patterns of political argumentation between arenas.
Our work addresses this research gap. First, we present a 17k-sentence corpus with annotation for argumentative passages (argument and their justifications, both their boundaries and their categories) across three German political arenas (plenary speeches, committee meetings, and press conferences), keeping the topic (COVID-19) constant. Our analysis of the corpus finds that contrary to expectations, justification by domain-specific expertise is more frequent in press conferences than in committee meetings. Second, we present a pilot study on automatically identifying such argumentative passages. The results show that boundaries are hard to pin down, and models predictions additionally suffer from confirmation bias.

## 1 Introduction

Even if existing democracies are often far from fulfilling the ideal of deliberative democracies (Bächtiger et al., 2018), political decision-making always contains elements of deliberation in parliaments, committees, and other arenas. In these fora, actors make claims and argue to support their positions and to attack opposing political actors. To advance their goal, they use various rhetorical and argumentation strategies, one of which is what Reyes has called "voices of expertise" (Reyes, 2011, 786), that is, the referral to experts or to expert knowledge in order to legitimise one's claim.

One should expect that argumentation strategies, in particular with regard to the appeal to expertise, are context dependent: Argumentation in plenary speeches should differ from argumentation in committees because the former address a general public, while the latter address mainly other MPs who are likely experts or at least experienced in the respective field. Consequently, one could expect that expert argumentation should be more common in committees than in plenary speeches. But studies so far have generally been limited to specific debates in one specific forum at a time. Comparative studies on argumentation strategies are very rare (e.g. Karlsson et al., 2024) because of the required amount of annotation.

In our study, we aim to make one step forward in the study of political debate across arenas by annotating, evaluating, and modeling arguments from three different corpora of political speech in Germany: plenary speeches, health committee meetings, and press conferences. We focus our analysis on the political discourse in these three arenas during the COVID pandemic in Germany (2020–2022). We choose this specific period and topic since during the COVID pandemic the reference to (medical) expert knowledge was omnipresent and we can therefore expect to see the use of expert arguments by many political actors, independent of whether they are themselves medical experts or not. Focusing on one domain also reduces the impact of topic shifts on our analysis.

Our paper makes two main contributions:

1. We present and analyse a manually annotated dataset of political discourse from three different German federal-level political arenas – to our knowledge, the first such corpus. It makes explicit argumentative passages, their justifications, and the type of justification used. Contrary to expectations, expert arguments play a bigger role in press conferences than in committee meetings.

2. We present a pilot study on the ability of

LLMs to identify argumentative passages in political arenas within and across fora. This would be a first step towards reducing manual effort for such analyses. We find, however, that argumentative passages are hard to identify: Many models are reluctant to conclude that documents are non-argumentative.

## 2 Related Work

**Argument Mining for Political Texts.** In computational linguistics, the automatic extraction of arguments ("argument mining") has grown rapidly over the last ten years, but has focused mainly on text types where ample data was available, including social media and student texts, see Cabrio and Villata (2018) for a review. Generalizing argument mining methods across text types is often challenging (Daxenberger et al., 2017; Schaefer et al., 2022). When it comes to political texts, most studies focus on uniform datasets of current interest, typically election campaigns (Lippi and Torroni, 2016; Visser et al., 2020). An exception is Poiaganova and Stede (2025) who evaluate various argument mining-related tasks within and across two political arenas (US presidential debates and UN security council speeches) and find reasonable generalization for most tasks across the two arenas. Although argument mining in parliamentary debates is already an established field of research, our approach is timely because it applies a zero-shot large language model to the task. Moreover, our annotation scheme extends existing political argumentation corpora by distinguishing different types of justification, including domain-specific justifications tailored to the analysis of expert argumentation.

**Political Argument Quality.** From a political science perspective, Steenbergen et al.'s Discourse Quality Index (DCI) is a normative evalutation of argumentation and in essence attempts to measure to which degree a debate fulfils the Habermasian ideal of deliberative politics (Steenbergen et al., 2003; Steffensmeier and Schenck-Hamlin, 2008). Research from a critical discourse analysis perspective puts a stronger focus on the use of various legitimization strategies, namely legitimization through emotions, through references to the future, through rationality, expertise, and altruism (Reyes, 2011). In the context of argument mining, it has been found that argument quality is best considered as a multidimensional phenomenon. Wachsmuth et al. (2017) have proposed a taxonomy for argumentation quality based on the cogency, reasonableness, and effectiveness of an argument.

**Political Arguments across Arenas.** Regarding arena specific argumentation strategies, political science research has highlighted differences between "debating" (e.g. UK) and "working" (e.g. Denmark, Sweden) parliaments, and between parliamentary "frontstage" and "committee backstage" (Karlsson et al., 2024) and pointed to more pragmatic argumentation in committees (Andone, 2016). On the computational side, there are few studies that compare properties of arguments across arenas. Reinig et al. (2024) characterize political speech in terms of speech acts, which are however distributed fairly complementarily to arguments. As mentioned above, Poiaganova and Stede (2025) consider two political arenas. Their mostly good generalization results imply that the arenas contain arguments with similar properties, but a deeper comparative analysis (e.g., with regard to expertise) was beyond the scope of their study. We included government press conferences as the third arena because they were important during the COVID-19 pandemic. They became a central venue for announcing new measures, with the health minister regularly appearing alongside leading scientific experts. framing the severity of the crisis and influencing subsequent media coverage (Hayek, 2024). While press conferences gained relevance during the pandemic, we expect a moderate share of expert argumentation.

## 3 Dataset

### 3.1 Corpus Selection

Our dataset of political discourse in the German political sphere consists of plenary speeches (Bundestagsreden), health committee protocols (Gesundheitsausschussprotokolle), and government press conferences (Bundespressekonferenzen). In this way, we aim to capture the full spectrum of parliamentary discursive arenas for one topical domain. Plenary speeches represent politicized front-stage deliberation and political competition. The health committee is a parliamentary back stage. And government press conferences are an interface between the general public and the parliament.

While plenary speeches are held on the agenda set in parliament and are delivered in monologue form, health committee and press conferences are dialogical formats, mostly organised as Q&A sessions. Press conferences usually begin with state-

ments by the participants, after which the government spokesperson responds to questions. The health committee is more or less scripted, as each party can invite (scientific) experts and question them according to the set agenda.

Our sampling strategy was determined in advance, as we sought to capture all COVID-related parliamentary discourse. Or period of observation starts at the beginning of 2020 with the onset of the pandemic and ends in April 2023 after all pandemic measures had ended and Health Minister Karl Lauterbach officially declared the pandemic over. We used the GermaParl Corpus (Blaette and Leonhardt, 2023) as data source for plenary speeches. The protocols of all public meetings of the health committee were downloaded from the archived websites of each legislative period (`https://www.bundestag.de/ausschuesse/gesundheit`) and the written transcripts of the Bundespressekonferenzen were obtained from `jungundnaiv.de`, a German journalist platform that provides complete coverage of these press conferences (Jung & Naiv, undated).

### 3.2 Preprocessing

The resulting dataset contains 118,745 speeches, 205 protocols and 512 press conferences. We then filtered the dataset for COVID-19-related content using a dictionary of 1,958 keywords, which was using terms from the Leibniz Institute for the German Language relating to the COVID-19 pandemic (Leibniz IDS, undated). We refined our approach with regular expression patterns to capture all relevant text segments and removed all terms not related to the medical domain. These are mostly terms from socially triggered discourse without medical import, such as 'click and collect', 'Drosten-Ultra' or 'Wiesn-Party'. This resulted in a total of 9,737 COVID-related texts (9,299 speeches, 159 protocols, 279 press conferences). We added metadata such as the speaker's first and last name, date, party, and session.

### 3.3 Annotation Schema

Following Nordin and Schiappa (2024, p. 8) we define our core concept, an *argument*, as a text segment consisting of one or more sentences and comprising a claim and its justification. Claims are statements that suggest an action, such as a demand, plan, position, or recommendation. Our definition thus represents a variant of the definition of arguments as claims supported by reasons.

Rather than adopting annotation schemes from the argument mining literature, our approach is grounded in the framework of practical argumentation and informal logic, particularly drawing on the Toulmin model. Initial attempts to operationalize speech act classification proved insufficient, as our first annotation sample did not align with the speech act categories. We therefore adopted a claim-based understanding of argumentation, showing how political claims are justified, rather than identifying argumentative units. While argument mining primarily focuses on the identification and the reconstruction of arguments, our aim is to capture modes of practical argumentation in political discourse. Specifically, we aim at classifying different modes of justification for political claims.

To further distinguish these different types of justifications, we used a manual annotation of a subset of texts as an inductive baseline to assess whether stable categories could be derived. Building on this foundation, we developed a classification scheme aimed at capturing systematic differences in argumentative styles. We sought to differentiate between expert-driven and layperson's argumentation. Therefore we defined three basic types: *evidence-based (disciplinary) justifications*, which rely on data, studies, models, or measurable evidence; *normative and pragmatic justifications*, which are based on values, norms, feasibility, or utility; and *ideological and analogical justifications*, which are based on opinions, rhetoric, or anecdotal and selective reasoning without verifiable evidence (Toulmin, 2003; Reyes, 2011; van Dijk, 2000). To capture expert argumentation, we assumed that only evidence-based (disciplinary) arguments qualify as expert arguments and added a binary indicator for whether the justification falls within the subject area. The annotation scheme was further extended to distinguish between medical and non-medical evidence-based justifications.

We sought to distinguish the COVID-19 discourse thematically in order to examine whether specific subtopics of the debate offer increased expert argumentation. We therefore classified each argument according to its subject. We used the manual annotation as a baseline and further refined it into six main categories: *healthcare* (treatment, access, personnel), *containment measures* (e.g. contact reduction, testing), *social issues* (socio-economic issues, education, rights), *research* (methods, data, expertise), international issues (cross-border issues), and *logistics* (implemen-

tation, administration, communication, funding).

Table 1 provides English translations of two argumentative passages found in the corpus together with their annotation. The examples show how the annotation scheme distinguishes between similarities and differences in argumentative reasoning. In both passages, a clear claim is supported by a justification, but the justifications differ. The first example represents an evidence-based argument, where the justification refers to an empirically testable mechanism. The second example, by contrast, exemplifies a normative argument, relying on a general appeal to urgency.

### 3.4 Annotation Procedure

The annotation guidelines were presented to two annotators, both political science students, and first tested on a small sample. This formed part of the inductive annotation procedure, in which new documents were selected and uploaded each week. To ensure a sufficient number of positive examples in our data, we pre-screen data for annotation using multiple factors. We sampled by date and political party, to cover different phases of the COVID-19 pandemic over time. We validated the selection by quantifying the presence of claims with an automatic classifier for detecting claims within German newspaper articles and manifestos (Blokker et al., 2020). This classifier provides a lower bound for the presence of claims, due to the shift in text type which tends to decrease recall. Our analysis showed that approximately 65% of the documents picked contained high number of claims, which we consider sufficient for our purposes.

The comparatively larger number of speeches is primarily due to structural differences. Speeches are substantially shorter and contain contributions by only a single speaker, whereas press conferences and committee protocols include multiple speakers and interaction sequences. The annotation effort per speech was approximately one quarter of the time required for the other documents.

The annotators coded the documents independently, while each document was annotated by both annotators, providing a reliable way to reduce subjectivity in the annotation process. Difficult cases were discussed after each annotation round to further refine the guidelines and assess their robustness. We used INCEpTION (Klie et al., 2018) as annotation tool and configured the platform in accordance with the annotation guidelines by defining the annotation layers (ARG/JUST), features (topic for ARG; justification type and domain for JUST), and corresponding tagsets (topics for ARG; types and domains for JUST). As a result, the guidelines were accessible during annotation, and also structurally reflected in the annotation environment. The estimated annotation speed was roughly 23–30 arguments per hour, when we consider 15-25 min per speech and one hour for the other documents due to their length. The full (German) guidelines are available in Appendix A.

After the first annotation round, we analyzed the inter-coder agreement (ICA) with the Gamma statistic (Mathet et al., 2015), a chance-corrected measure of agreement between sequence annotations that incorporates both labeling agreement (like Cohen's $\kappa$) and segmentation choice (which $\kappa$ does not take into account). Like for $\kappa$, its range is $[-1, 1]$, where 1 indicates perfect agreement and 0 agreement at chance level.

As the second-to-last row of Table 2 shows, ICA after initial annotation was mediocre, mainly due to differences in the identification of argument boundaries. We therefore created a gold standard in iterative steps, similar to other semantic annotation projects (e.g., OntoNotes, Hovy et al. 2006). First, matching annotations were merged in the INCEpTION tool. Second, arguments and justifications without matching codes were manually reviewed by a third annotator with advanced expertise in political science and refined in accordance with the codebook. Thereafter, the ICA between the created gold standard and the initial annotations was calculated (final row of Table 2). This resulted in $\gamma$ scores around 0.7, reflecting the success of the adjudication procedure and demonstrating that the gold standard indeed represents a consensus between the two annotations. Still, we take away that it is hard to get annotators to agree perfectly on argument spans.

We do not carry out more detailed analysis of category correspondences, since these are hard to align in a combined segmentation-and-categorization task.

### 3.5 Corpus Findings

Our corpus consists of 9,299 parliamentary speeches, 159 committee protocols, and 279 press conferences, for a total of 9,737 texts. Note that the two latter types collapse multiple contributions in one document and are therefore much longer. Of these, we fully annotated 87 documents, with an average of 19.47 argument spans per documents.

| Evidence-based Argument | Normative Argument | Ideological Argument |
|---|---|---|
| *"Strict contact restrictions are necessary because they can halt and reverse the spread of infection."* | *"Something needs to be done. If we ignore the problem now, the situation will get worse."* | *"It seems to be noone from Ms. Merkel, Mr. Spahn and Mr. Altmaier knows - because they are childless - what they are demanding from our youth. It is a catastrophe."* |
| **Claim:** "Strict contact restrictions are necessary [...]" | **Claim:** "Something needs to be done." | **Claim:** "It is a catastrophe." |
| **Topic:** Containment measures | **Topic:** Unspecified | **Topic:** Social |
| **Justification:** "[...] because they can halt and reverse the spread of infection." | **Justification:** "If we ignore the problem now, the situation will get worse." | **Justification:** "[...]because they are childless[...]" |
| **Type:** Evidence-based | **Type:** Normative/pragmatic | **Type:** Ideological/analogical |
| **Medical:** Yes | **Medical:** No | **Medical:** No |

Table 1: Agumentative spans and annotated labels.

| | BT | GA | BPK | Total |
|---|---|---|---|---|
| Documents | 60 | 15 | 12 | 87 |
| Contributions | 60 | 1046 | 1450 | 2465 |
| Sentences | 1881 | 9350 | 5888 | 17119 |
| Avg. Sent./Text | 31.35 | 623.33 | 490.67 | 196.77 |
| ARG Spans | 409 | 1079 | 201 | 1689 |
| ARG / 100 Sent. | 21.74 | 11.54 | 3.41 | 9.87 |
| JUST Spans | 457 | 1245 | 216 | 1918 |
| JUST / 100 Sent. | 24.30 | 13.32 | 3.67 | 11.20 |
| Initial ICA ($\gamma$) | 0.46 | 0.37 | 0.39 | |
| Avg. ICA with Gold ($\gamma$) | 0.72 | 0.69 | 0.73 | |

Table 2: Descriptive statistics by arena: BT (Bundestag speeches), GA (Gesundheitsausschuss protocols), BPK (Bundespressekonferenz press conference)

Table 2 shows statistics for the resulting annotated corpus, both in total and broken down by arena. We identified 1,689 argument spans (ARG) and 1,918 justification spans (JUST), amounting to 3,607 spans overall. On average, this corresponds to 18.35 ARG spans per 100 sentences and 20.65 JUST spans per 100 sentences. The ARG spans were differentiated by topic, while the JUST spans were differentiated by justification type.

Looking at the topic distribution, logistics accounts for 60.1%, followed by healthcare at 15.1%. Social issues amount to 10.6%, and containment measures to 8.3%, while international issues account for 2.8% and research for 2.8%.

Regarding justification types, evidence-based justifications make up only 7.1% of all JUST spans, while normative and pragmatic justifications account for 88.4% and ideological/analogical justifications for 4.4%. Among the evidence-based JUST spans, 61.5% were classified as medical and 38.5% as non-medical. The sources also differ in their argument density. Speeches show the highest average, with 23.02 ARG and JUST spans per text, followed by protocols (12.43). Press conferences contain only 3.54 ARG and JUST spans per text.

Overall, most JUST spans appear to be normatively driven justifications. However, when looking at the share of evidence-based justifications, BPK shows the highest proportion at 16.7%. Speeches account for 6.6% and protocols for 5.7%.

The data reveal a discipline-specific concentration of evidence-based arguments, which is in line with our assumption that expert argumentation primarily takes place in evidence-based arguments. The observed argument density also corresponds to our expectations: Although plenary speeches are a central site of argumentation, the Q&A format of press conferences appears to constrain argument density, with committee protocols falling between the other two sources. The press conferences appear to be the primary venue for expert argumentation, while plenary speeches are driven by normative justification. Contrary to expectations, the health committee is only third.

These findings align directly with our expectation that appeals to expertise are context dependent but contrast with the expectation that such forms would be concentrated in committees, where argument density is high and expert knowledge might be expected most strongly. Although previous research focuses on public "frontstage" debate versus expertise-oriented "backstage" committee deliberation, we observe that expert argument appears more often in government press conferences, rather than in committee deliberation (Karlsson et al., 2024; Andone, 2016). Although committees show a dense argumentative interaction, expertise-based justification appears most prominently in press conferences. In plenary speeches, we observe more normative reasoning: Justifications rest on general appeals to collective protection rather than on empirical evidence. For example, one MP

argues that restrictions on fundamental rights are necessary to enable containment measures aimed at protecting the population.[1]

The prominence of expert arguments in government press conferences may reflect their institutional logic: Arguments in press conferences are used to communicate governmental decisions in a formalized setting, where speakers usually first deliver a relatively short statement and then respond to the journalists' questions. This setting seems to trigger expert-like justifications to provide reasonable and deliberative answers. Press conferences thus tend to feature dense and evidence-based reasoning, where statements e.g. about increasing incidence rates, the spread of the B.1.1.7 variant, and increasing ICU occupancy are used to justify concrete government decisions, reflecting a stronger reliance on empirically grounded argumentation.[2]

In the health committee, experts provide descriptive input, e.g., when discussing staff vaccination rates. This leads to a high number of normative justifications. We attribute this to the interaction between politicians and invited experts, where justifications are more normative than evidence-based.[3]

Thus, we can see overall that expertise is not simply tied to institutional proximity to expert knowledge, but also to communicative function and audience. At the same time, the more strongly normative plenary speech aligns with theories of deliberative and representative politics that emphasize public justification, value contestation, and political positioning in plenary arenas (Steenbergen et al., 2003).

[1]Example: "Ja, zum Schutz der Bürger vor Corona sind Grundrechtseingriffe erforderlich, etwa durch Hygienekonzepte oder Abstandsmaßnahmen."

[2]Example: "Es gibt das Auftreten neuer Virusvarianten, das die epidemiologische Lage verändert hat. Es gibt insbesondere die deutlich ansteckendere Virusvariante B.1.1.7. Es gibt stetig steigende Inzidenzen. Wir liegen heute bei einem Wert von 140,9. Wir haben eben auch eine wieder stetig steigende Zahl von Menschen, die auf den Intensivstationen behandelt werden müssen. Deswegen hat das Kabinett heute eine Ergänzung des Infektionsschutzgesetzes beschlossen, und zwar, wie ich gesagt habe, als Formulierungshilfe für die Fraktionen von CDU/CSU und SPD."

[3]Example: "Ich habe im Moment bei den Mitarbeitern eine Impfquote von 98 Prozent. Zu keinem Zeitpunkt hatten wir größere Probleme, außer den üblichen, den Dienstplan zu gestalten. In Moment habe ich, wenn ich die gesamte Gruppe anschaue, sechs Stellen zu besetzen. Das sind also auch hier keine Horrorszenarien, wie sie immer wieder kommuniziert werden."

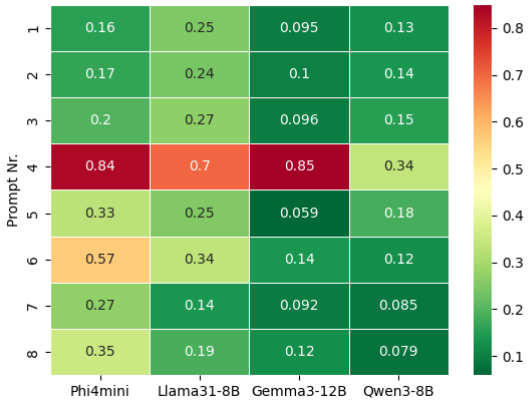


Figure 1: Return rates of unfaithful quotations for combinations of models and prompt templates.

### 3.6 Text Preparation

For modeling use, we split the datasets individually into three parts: training, development, and test. We decided to make these parts equal-sized (33% each), to accommodate the various possible approaches to modeling (traditional training/evaluation; pre-training and fine-tuning; zero-shot/few-shot LLM use) while keeping the test set large enough for robust evaluation.[4]

## 4 Pilot Modeling Study

We now present the results of a pilot study limited to the automatic recognition of argumentative spans for the different arenas in our annotated corpus. From our perspective, this task is the first step in a future more comprehensive modeling endeavour that will also consider other aspects of the annotation that we have creating, including claim and justification spans, claim domains and justification types. We restrict ourselves to the argument recognition subtask since (a) it is the logically first step when facing unanalyzed political text: all other steps presuppose recognized arguments; (b) a reliable model for just this step would already play an enabling role in the analysis of argumentation in political texts across arenas (cf. Section 1).

### 4.1 Model Choice

Since the models need to process German data, we require LLMs that have a good German proficiency. Monolingual German models are rare, and those that exist such as LLäMmlein (Pfister et al., 2025)

[4]The dataset is available at Zenodo: https://doi.org/10.5281/zenodo.21718040

have undergone no or only basic instruction tuning. Given that our task is complex, we focus on multilingual instruction-tuned models. For reasons of practicality and reproducibility, we restricted ourselves to open-weights models in the range between 4B and 12B parameters (Liesenfeld et al., 2023). Specifically, we investigated Phi4-mini (4B, Abouelenin et al. 2025), Llama3.1 (8B, Grattafiori et al. 2024), Gemma3 (12B, Kamath et al. 2025) and Qwen3 (8B, Yang et al. 2025).

### 4.2 Experimental design

We follow the current standard usage mode for zero-shot text processing with large language models: we craft a prompt for our task, present the prompt with each input to LLMs, and parse the outputs.

**Prompt selection.** The model was instructed to return argumentative passages related to COVID from a given text. This is a task with a fairly complex output structure, namely a list of argumentative passages. Models can misquote the input, justify their output (contrary to instructions), or add arbitrary comments. Since we considered it unlikely that different LLMs would all work well with the same single prompt, we formulated a set of 8 prompt candidates, listed in Appendix B, that varied in length and detail of definition, closeness to guidelines, instructions on output form as well as language of instruction. These templates were evaluated on the train set. We then selected the prompt that led to the lowest number of unfaithful quotations from the input, across arenas. The results are shown in Figure 1. Note that this evaluation does not require gold standard annotation, only input text. The results show that one prompt (#4) fails for all models and that models indeed prefer different prompts (e.g., Phi4mini prompt #1 vs. prompt #8 for Qwen3-8B). For the respective best prompt, the unfaithful quotation rate lies between 6% (Gemma3) and 16% (Phi4mini).

**Answer interpretation.** Our answer interpretation component accommodates variability in LLM output by trying to identify three types: (a), a list of arguments with quotation marks; (b), any other kind of list of arguments; (c) an empty list, typically accompanied with a justification. For (a), the text in-between quotation marks is extracted. For (b), we match a list with a regular expression. When some bullet point is recognized, everything located on the line is considered an argument. For (c), we look for keyphrases related to an empty return, such as *"leere Liste"* or *"keine argumentativen Textstellen"*. Outputs that remain unrecognized are treated as an empty list for the purposes of evaluation; however we keep track of the number of such 'invalid response' cases and report them separately.

### 4.3 Evaluation Metrics

We assess model predictions with a couple of evaluation metrics. First, we consider three variants of sequence-level $F_1$ score. All of these are computed at the level of annotation spans (not individual tokens) and differ in what they consider true positives (TPs): In the strict match condition, a model-predicted span is a TP only if it exactly matches a gold standard TP. In the inclusion condition, a model-predicted span is a TP if it is included in a gold standard span. In the partial match condition, each predicted span is aligned with the best-aligned gold span, and the TP count is the fraction of their overlap divided by predicted span length. Thus, strict match is a harsh criterion while partial match rewards models already for mostly wrong predictions. Inclusion attempts to strike a balance by rewarding models for not overpredicting spans.

The other metric we consider is Gamma ($\gamma$), which is usually used as a chance-corrected metric for inter-coder agreement (ICA) in sequence annotation (cf. Section 3.4). Agreement between model-predicted spans and gold-standard spans with $\gamma$ can thus be compared to ICA numbers from above.

Finally, we report the percentage of invalid model responses, the number of non-argumentative contributions according to models and gold standard, and predicted argument density, measured as number of arguments per 100 sentences.

### 4.4 Baselines

As points of comparison for the LLM-based approach, we consider three simple baselines. Baseline 1 does not predict any arguments. Since this baseline has a recall of 0, its performance is always zero. Baseline 2 predicts that the whole text is a single argument. Since this baseline has a very low precision, its F1 score is also very low and we do not report detailed results. Baseline 3 predicts that every sentence is (its own) argument. This is the main baseline worth comparing against.

### 4.5 Modeling Results

The results for extracting argumentative passages from contributions in the three arenas are shown in

| | | Baseline 3 | Phi4mini | Llama3.1-8B | Gemma3-12B | QWen3-8B |
|---|---|---|---|---|---|---|
| BT | Pr/Re/F1(strict) | 0.04 0.20 0.07 | 0.09 0.06 0.07 | 0.05 0.02 0.03 | 0.15 0.07 **0.10** | 0.13 0.04 0.07 |
| | Pr/Re/F1(inclusion) | 0.62 0.77 **0.69** | 0.70 0.33 **0.44** | 0.60 0.22 0.32 | 0.67 0.26 0.38 | 0.70 0.20 0.31 |
| | Pr/Re/F1(partial) | 0.67 0.79 **0.73** | 0.78 0.35 **0.49** | 0.73 0.25 0.37 | 0.79 0.29 0.43 | 0.80 0.22 0.35 |
| | Gamma (ICA) | -0.12 | 0.09 | 0.01 | 0.02 | **0.14** |
| | invalid responses | | 0/20 | 0/20 | 0/20 | 0/20 |
| | #pred/gold empty list | 0/0 | 0/0 | 0/0 | 0/0 | 2/0 |
| | avg. args/100 sent. | 100 | 14.4 | 10 | 10.8 | 8.1 |
| GA | Pr/Re/F1(strict) | 0.01 0.09 0.02 | 0.05 0.07 0.06 | 0.06 0.07 0.06 | 0.06 0.10 **0.07** | 0.11 0.04 0.06 |
| | Pr/Re/F1(inclusion) | 0.51 0.80 **0.62** | 0.66 0.51 **0.57** | 0.78 0.47 0.56 | 0.57 0.52 0.54 | 0.62 0.19 0.29 |
| | Pr/Re/F1(partial) | 0.53 0.81 **0.64** | 0.69 0.52 **0.59** | 0.74 0.48 0.58 | 0.60 0.53 0.57 | 0.71 0.21 0.33 |
| | Gamma (ICA) | -0.09 | 0.33 | **0.35** | 0.22 | 0.33 |
| | invalid responses | | 14/350 | 0/350 | 0/350 | 0/350 |
| | #pred/gold empty list | 0/176 | 171/176 | 175/176 | 67/176 | 235/176 |
| | avg. args/100 sent. | 100 | 17.6 | 14 | 21 | 4.5 |
| BPK | Pr/Re/F1(strict) | 0.00 0.07 0.01 | 0.03 0.09 0.04 | 0.07 0.06 **0.07** | 0.01 0.04 0.01 | 0.07 0.07 **0.07** |
| | Pr/Re/F1(inclusion) | 0.14 0.73 0.23 | 0.19 0.39 **0.25** | 0.15 0.32 0.20 | 0.13 0.48 0.20 | 0.25 0.23 0.24 |
| | Pr/Re/F1(partial) | 0.15 0.75 0.25 | 0.20 0.41 **0.27** | 0.17 0.34 0.22 | 0.14 0.50 0.21 | 0.26 0.24 0.25 |
| | Gamma (ICA) | 0.17 | 0.43 | 0.42 | 0.28 | **0.44** |
| | invalid responses | | 56/490 | 16/490 | 0/490 | 3/490 |
| | #pred/gold empty list | 0/406 | 363/406 | 331/406 | 164/406 | 379/406 |
| | avg. args/100 sent. | 100 | 11.6 | 10.7 | 25.8 | 4.2 |

Table 3: Performance of LLMs on argumentative passage extraction in three arenas (BT, Bundestag; GA, Gesundheitsausschuss; BPK, Bundespressekonferenz). Highest results for each metric in each arena boldfaced. Baseline 3: Every sentence is its own argument.

Table 3. We use the best prompt for each model (cf. Figure 1). Our main observations are as follows.

**Exact match $F_1$.** Since exact match is a harsh success criterion for the models, results are in the single digits for almost all models. This is not surprising, given the difficulty of human annotators to agree on exact boundaries (cf. Section 3.4) but underlines the difficulty to precisely delineate argumentative passages in our data. It is Gemma3, the largest model, which obtains the highest numbers on Bundestag and Gesundheitsausschuss.

**Partial / inclusion $F_1$.** According to these more lenient metrics, the identification of argumentative passages is not impossible, but still a seriously challenging task. Phi4mini is the overall winner, showing robust performance across all three arenas. This is striking, given that it is the smallest among our LLMs (4B parameters). Gemma3 and Llama3.1 follow in second place for Bundestag and Gesundheitsausschuss, while QWen shows rather good performance on Bundespressekonferenz – see below for an explanation.

**Gamma scores.** The Gamma scores indicate that some of the models perform close to or even at chance level, notably Llama3.1 and Gemma3 on Bundestag. This is different for the two other arenas, where Gamma scores reach 0.35 (Gesundheitsausschuss) and 0.44 (Bundespressekonferenz).This is a level comparable to the agreement among human annotators (cf. Table 2). According this metric, QWen3 is the model with the overall most robust performance.

**Argument density.** The results above can be explained, to an extent, by the differences in argument density among arenas. According to Table 2, the gold standard argument densities range between 22 arguments per 100 sentences (Bundestag), 12 arguments per 100 sentences (Gesundheitsausschuss) and 3 arguments per 100 sentences (Bundespressekonferenz). We observe striking differences among models regarding the predicted argument density. All models underpredict arguments on the Bundestag data (8–14 arg. / 100 sent.) but most mildly overpredict arguments on the Gesundheitausschuss and extremely overpredict on the Bundespressekonferenz (except QWen3). The most extreme case is Gemma3 which predicts 26 arguments per 100 sentences on BPK, almost 10 times the correct density. This naturally results in low precision. In comparison, Phi4mini estimates the argument density most accurately on Bundestag and Gesundheitsausschuss, which accounts for its good performance on these corpora. Only on Bun-

despressekonferenz, its density estimate is off; here, it is QWen3 which comes closest.

The differences between assumed argument density across arenas is particularly striking since all of these models are zero-shot. The only contact that the models have had with the evaluation data is the prompt selection step (cf. Section 4.2).

Across all models and arenas, the closer a model manages to get to the argument density in the gold standard, the better its predictions: F1 scores (partial match) and density accuracy (difference between gold and predicted density, divided by gold density) are inversely correlated (Spearman's $\rho$=-0.56, $n$=12, $p$=0.06). Arguably, the lower argument density of Bundespressekonferenz stems from the fact that about 80 percent (406/490) of the contributions are purely descriptive, with no argumentation present. As the numbers on non-argumentative contributions according to models and gold-standards show, this is a major stumbling block for some of the models.

**Baseline.** The baseline which assumes that every sentence is a separate argument predicts, by definition, an argument density of 100, and never returns an empty argument list. It performs surprisingly strongly. In terms of F1 scores, the baseline outperforms the best LLMs on the Bundestag and Gesundheitsausschuss arenas and does about as well as the median LLM on Bundespressekonferenz. However, the much smaller Gamma scores on all arenas (even negative on BT and GA) provide evidence for the essentially random nature of the predictions. Again, the differences among the arenas reflect their differences in argument density: a baseline which assumes that every sentence is an argument performs better for arenas with denser argumentation.

## 5 Conclusions and Future Work

This paper investigated the differences in argumentation patterns among different arenas in German political discourse during the COVID-19 pandemic. In the absence of existing corpora to ground our analysis, we make two contributions: (a) manual annotation of a 17k-sentence corpus comprising three important political arenas with argumentative passages, justifications, and corresponding categories; and (b) a pilot study to automatically recognize argument boundaries as a first step towards automating the information that we annotated.

Our main insights are as follows. First, identifying the boundaries of argumentative passages in political discourse is difficult both for human annotators and for models. We also found lower agreement compared to previous studies (Haddadan et al., 2019; Poiaganova and Stede, 2025). The extent to which this is due to our formulation of the task as sequence identification (and not sentence-level classification) is a matter of future work. An alternative route would be to embrace the differences among annotators in the spirit of perspectivism (Falk et al., 2024).

Second, we found major differences in argumentation patterns among our three arenas: the Bundespressekonferenz, while having the lowest argument density, shows the highest proportion of domain expertise-related justifications, while in plenary speeches normative justifications dominate.

Third, a major contributing reason for the difficulty that zero-shot LLMs have in identifying argumentative passages is that they struggle with estimating the argument density in input texts. This is likely because the models are trained to accept the presuppositions of instructions (in our case, the presence of arguments in the documents) and thus are reluctant to dismiss these presuppositions. As a consequence, our LLMs found it hard to beat a simple 'everything is an argument' baseline on the Bundestag and Gesundheitsausschuss texts. In future work, we plan to explore the role of simple embedding-based classifiers to 'pre-screen' documents for argumentative content (Ruiz-Dolz and Lawrence, 2023).

Other modeling options that were out of scope for this first study is the use of few-shot setups which could provide LLMs with priors about argument density (Schaefer et al., 2022); replacing LLMs with 'more traditional' embedding-based classifiers that have shown good performance for argument recognition in previous work (Poiaganova and Stede, 2025); and modeling the remaining aspects of our annotation (justification spans as well as argument and justification categories).

## Limitations

At the corpus level, our annotation effort is focused on a single domain (COVID-19). While we believe this to be a reasonable choice (cf. Section 3.1), we acknowledge that our findings may not generalize straightforwardly to other domains. Related to this, we only consider a single committee (the health committee – Gesundheitsausschuss); investigations

of other domains should consider other committees.

At the model level, we have considered five 'consumer-sizes' open-weights LLMs but considered neither simpler solutions (embedding-based models) nor very large properietary LLMs which might provide considerably better performance. Prompts were also selected from a small number of templates.

## References


Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, and 55 others. 2025. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *Preprint*, arXiv:2503.01743.

Corina Andone. 2016. Argumentative patterns in the political domain: The case of European parliamentary committees of inquiry. *Argumentation*, 30(1):45–60.

Andreas Blaette and Christoph Leonhardt. 2023. GermaParl corpus of plenary protocols. Technical report, zenodo. [Data set].

Nico Blokker, Erenay Dayanik, Gabriella Lapesa, and Sebastian Padó. 2020. Swimming with the tide? positional claim detection across political text types. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 24–34, Online. Association for Computational Linguistics.

André Bächtiger, John S. Dryzek, Jane Mansbridge, and Mark E. Warren. 2018. *Deliberative Democracy. An Introduction*, page 1–31. Oxford Handbooks. Oxford University Press, Oxford.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: A data-driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018)*, pages 5427–5433.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? Cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Neele Falk, Andreas Waldis, and Iryna Gurevych. 2024. Overview of PerspectiveArg2024 the first shared task on perspective argument retrieval. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 130–149, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Lore Hayek. 2024. Media framing of government crisis communication during Covid-19. *Media and Communication*, 12:7774.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Jung & Naiv. undated. Bundespressekonferenz transcripts. Accessed 25.06.2025.

Kishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 196 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Christer Karlsson, Thomas Persson, and Moa Mårtensson. 2024. Do members of parliament express more opposition in the plenary than in the committee? Comparing frontstage and backstage behaviour in five national parliaments. *Parliamentary Affairs*, 77(1):173–195.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, New Mexico, USA.

Leibniz IDS. undated. Word list for the COVID-19 pandemic. Accessed 25.06.2025.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, page 1–6. Association for Computing Machinery.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma ($\gamma$) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

John P. Nordin and Edward Schiappa. 2024. *Argumentation: Keeping Faith with Reason*, 2 edition. Routledge, New York.

Jan Pfister, Julia Wunderle, and Andreas Hotho. 2025. LLäMmlein: Transparent, compact and competitive German-only language models from scratch. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2227–2246, Vienna, Austria. Association for Computational Linguistics.

Maria Poiaganova and Manfred Stede. 2025. From debates to diplomacy: Argument mining across political registers. In *Proceedings of the 12th Argument mining Workshop*, pages 205–216, Vienna, Austria. Association for Computational Linguistics.

Ines Reinig, Ines Rehbein, and Simone Paolo Ponzetto. 2024. How to do politics with words: Investigating speech acts in parliamentary debates. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, page 8287–8300, Torino, Italia. ELRA and ICCL.

Antonio Reyes. 2011. Strategies of legitimization in political discourse: From words to actions. *Discourse & Society*, 22(6):781–807.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*, pages 1–10, Singapore. Association for Computational Linguistics.

Robin Schaefer, René Knaebel, and Manfred Stede. 2022. On selecting training corpora for cross-domain claim detection. In *Proceedings of the 9th Workshop on Argument Mining*, pages 181–186, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Marco R. Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1(1):21–48.

Timothy Steffensmeier and William Schenck-Hamlin. 2008. Argument quality in public deliberations. *Argumentation and Advocacy*, 45(1):21–36.

Stephen E. Toulmin. 2003. *The Uses of Argument*, 2 edition. Cambridge University Press.

Teun van Dijk. 2000. *Ideology: A Multidisciplinary Approach*. SAGE Publications Ltd, London.

J. Visser, B. Konat, R. Duthie, M. Koszowy, K. Budzynska, and C. Reed. 2020. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54:123–154.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, page 176–187, Valencia, Spain. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

## A Full Annotation Guidelines

(follow on next complete page)

# Annotationsrichtlinien – Argumentation

Es sollten Argumente in den gegebenen Texten annotiert werden.
Ein Argument kann **ein oder mehrere Sätze** enthalten. Ein Argument besteht aus Begründungen und einer Behauptung, die begründet wird.

Schritte zur Annotation:
**1.** Finde eine Behauptung im Text
(→ Definition von Argumenten)
**2.** Kennzeichne den Text um die Behauptung (ca. 2-3 Sätze)
**2. 1. Weise ein Thema** dem Abschnitt **zu**
(→ Annotation von Themen)
**2.2. Kennzeichne die Begründungen**, die innerhalb dieses Textabschnitts enthalten sind
(→ Definition von Argumenten)
**2.2.1** Gebe bei jeder Begründung an, **ob** es sich um **eine disziplinäre Begründung** handelt
(→ Annotation von Argumenten)
2.2.1.1 **wenn Begründung** (Evidenz-theoretisch), dann weitere, binäre Unterscheidung in medizinisch oder nicht medizinisches Begründung

Bemerkungen:
• Fokus auf **eindeutige Abschnitte**
• unsichere Annotation vermeiden, dennoch sollen bei impliziter Forderung oder Schlussfolgerung das Argument annotiert werden, solange die implizite Forderung erkannt wird und das Annotationsschema funktioniert, d.h. **2.1.** und **2.2.** können angewandt werden
• Eine Schlussfolgerung in einem Argument darf eine Begründung in einem anderen sein

# Beispiel

Meine sehr verehrten Damen und Herren, die Fraktion der FDP vertritt in ihrem Antrag die Au fassung, dass es in der Bevölkerung eine Unsicherheit über Teststrategien auf das Corona-Virus gebe und sich diese zuletzt weiter vergrößert habe. Es müsse für die Menschen nachvollziehbar sein, nach welchen Kriterien die begrenzten Test-Ressourcen verwendet würden. Daher müsse die nationale Teststrategie an den Prinzipien der Zielgerichtetheit und des Risikogruppenschutzes ausgerichtet werden. Die Strategie sowie die Quarantäne und Testempfehlungen müssten stetig anhand der wissenscha tlich aktualisierten Infektionsdynamik und Viruslast angepasst werden, da sich die wissenscha tliche Datenbasis zu SARS-CoV2 ständig erweitere. Im Einzelnen fordert die Fraktion unter anderem, Altenp legekrä te alle 14 Tage zu testen, um Risikogruppen zu schützen, und Patienten mit Symptomen weiterhin priorisiert zu testen. Außerdem solle dafür Sorge getragen werden, dass es mit dem Ausbau digitaler Meldewege und dem 24/7-Betrieb von Laboren und Gesundheitsämtern zu deutlich weniger Verzögerungen der Ergebnisübermittlungen und eventueller Maßnahmen komme. Ich freue mich auf eine angeregte Diskussion mit unseren Sachverständigen zu diesen Fragen.

Argument 1:
Es müsse für die Menschen nachvollziehbar sein, nach welchen Kriterien die begrenzten Test-Ressourcen verwendet würden. Daher müsse die nationale Teststrategie an den Prinzipien der Zielgerichtetheit und des Risikogruppenschutzes ausgerichtet werden.

Argument 2:
Die Strategie sowie die Quarantäne und Testempfehlungen müssten stetig anhand der wissenscha tlich aktualisierten Infektionsdynamik und Viruslast angepasst werden, da sich die wissenscha tliche Datenbasis zu SARS-CoV2 ständig erweitere.

Argument 3:
Im Einzelnen fordert die Fraktion unter anderem, Altenp legekrä te alle 14 Tage zu testen, um Risikogruppen zu schützen, und Patienten mit Symptomen weiterhin priorisiert zu testen.

Argument 4:
Außerdem solle dafür Sorge getragen werden, dass es mit dem Ausbau digitaler Meldewege und dem 24/7-Betrieb von Laboren und Gesundheitsämtern zu deutlich weniger Verzögerungen der Ergebnisübermittlungen und eventueller Maßnahmen komme.

– Behauptung - Begründung

# Definition von Argumenten

Argumentation:
besteht aus einer **Behauptung** und **Begründungen**, wobei nicht immer beide Elemente explizit vorhanden sind. Die Begründungen sollen gekennzeichnet werden. In schriftlichen Quellen können Argumente impliziter, bspw. eher als Schlussfolgerung/Erklärung auftreten.

**Behauptung:**
Hier wird auf Aussagen geachtet, die eine Handlung suggerieren.

z.B. **Aufforderung** („Das soll verboten werden.“, „Ich fordere, dass mehr dazu gemacht wird.“)
**Vorhaben** („Wir machen so weiter.“, „Es ist wichtig, das es besprochen wird.“)
**Standpunkt** („Der Verein ist dagegen.“, „Ich schließe mich an.“)
**Empfehlung/Einschätzung** („Unserer Meinung nach, muss das gelöst werden.“, „Das soll aufgegriffen werden.“)

oder **Schlussfolgerung**, …

**Begründungen** können in eine (oder mehrere) Kategorien fallen:

**Fakten und Bewertungen**

z.B. „Die Infektionszahlen sind gefallen. Die Maßnahmen waren also effektiv. (Deswegen sollen wir so weitermachen.)“
„Aufgrund der steigenden Tendenz, sollten die Maßnahmen angepasst werden.“

**bestehende Mängel/Probleme/Schwierigkeiten**
und **Abwertung** anderer Vorschläge

z.B. „Derzeit haben wir nicht die Kapazitäten. (Wir wollen das lösen.)“
„Die vorherige Aussage stimmt so nicht. Das wird so nicht funktionieren. (Wir sollen es nicht so umsetzen.)“

**Absicht/Ziel**

z.B. „(Es soll online angeboten werden,) damit alle Interessierten teilnehmen können.“
„Wir müssen den Menschen Sicherheit geben. (Daher verbessern wir die Transparenz.)“

**Prognosen/Aussicht/Vorteile und (unbeabsichtigte) Folgen**

z.B. „Wenn wir das Problem jetzt ignorieren, wird die Lage sich verschlechtern.“
„Arbeit im HomeOffice verringert die Ansteckungsgefahr und auch der Arbeitsweg fällt weg.“

**Veranschaulichungen**

z.B. „Derzeit ist das Amt in (Bezirk) zu 90% ausgelastet. (Wir müssen mehr Angebot schaffen.)“
„Angenommen eine Person in der Bahn ist infiziert. Dann kann sich das Virus in dem Raum sehr schnell verbreiten. (Deswegen sollen wir das Problem angehen.)“

**(Personen)verweise**

z.B. „Wie mein Vorredner schon gesagt hat, (…)“

**Bezug auf einen "offenkundigen" Grund**
kein Inhalt, bezieht sich auf "gesunden Menschenverstand" oder "allgemein vorhandenes Wissen"

z.B. "Es ist offensichtlich, (etwas soll getan werden)."
"Wie man es kennt, (...)"

# Annotation von Themen

Zu jedem **Textabschnitt** soll **ein Thema** aus der folgenden Liste zugewiesen werden. Bei Uneindeutigkeit kann auf Themensubjekt der **Begründung** geschaut werden.

## Gesundheitsversorgung

Umfasst ausschließlich medizinische Versorgung und Behandlungsmethoden, inkl. deren Verfügbarkeit, Zugang und Personal.

- Personal
  *Krankenhäuser / Pflege / Kitas*
- Impfstoff
  *Wirkung / Kampagne*
- Public Health
  *Gesundheitliche Folgen / Risikogruppen / Long Covid*

## Eindämmungsmaßnahmen

Aussagen über Maßnahmen zur Reduzierung von Infektionen, wie z.B. Kontaktreduzierung und Testmöglichkeiten.

- Kontaktreduzierung
- Maßnahmen (z.B. Ausgangssperren)
  *Prävention*
- Impfen als Maßnahme
  *Quote*
- Test
  *Prozess / Genauigkeit / Zugang*

## Soziales

Gesellschaftliche Themen, die den Alltag der Menschen betreffen.
Darunter fallen beispielsweise sozio-ökonomische sowie psychosoziale Probleme, Selbstbestimmungsrechte oder auch das (Aus-)Bildungsangebot.

- Auswirkungen und Probleme
  *Sozioökonomische oder psychosoziale*
- Ausbildung
- Solidarität
- Selbstbestimmung
  *Grundrechte*
- Impfen als soziales Thema

## Forschung

Aussagen über Förderung der Forschung und deren Integration in Maßnahmen.
Es kann dabei genauer auf die Methodiken eingegangen werden.

- Einbeziehung / Kooperation mit Experten
- Datenerhebung / Datenmangel
- Behandlung / Krankheitsverlauf / Behandlungsmethoden / Diagnostik
- Prognosen
- Impfstoff
  *Entwicklung*

## Internationales

Länderübergreifende Themen, die die Situation, Menschen und Regelungen auch außerhalb von Deutschland ansprechen.

- Internationale Gesundheit
- Wirtschaft
- Einreise
- Naturschutz und Nachhaltigkeit

## Logistik

Themen, die sich mit konkreter Umsetzung von Vorhaben befassen.
Das umschließt z.B. Verwaltung, Digitalisierung von Prozessen, Finanzierung und Verwaltung des Informationsflusses.

- Finanzierung / Kapazitäten / Verwaltung / Umsetzung
- Informationsfluss
  *Transparenz / Kommunikation / Aufklärung*
- Kontrollen / Sicherheit
- Digitalisierung
- Impfstoff
  *Beschaffung*

# Annotation von Argumenten

Jeder **Begründung** soll **eine Art** zugewiesen werden:

## Evidenz-theoretisch (disziplinär/wissenschaftlich)

Stützt sich, Erhebungen, Experimente oder Statistiken. Leitet Aussagen aus bestehenden, jeweils aus den disziplinarischen Modellen, Theorien oder konzeptionellen Rahmen her. Der Verweis auf eine Studie reicht aus, um die Kategorie zu erfüllen, da wir annehmen, dass das nicht "missbraucht" wird.

**Adjektive:** messbar, überprüfbar, evidenz-/zahlengestützt, konsistent, sachlich, stringent

**Beispiele:**

„Aufgrund der steigenden Tendenz sollten die Maßnahmen angepasst werden."

„Konsequente Kontaktbeschränkungen sind notwendig, weil damit das Infektionsgeschehen gestoppt und umgekehrt werden kann."

## Ideologisch und analogisierend

Behauptungen beruhen weitgehend auf anekdotischen Beobachtungen, Meinungen, Rhetorik oder selektiver Darstellung ohne überprüfbare Belege. Appellative und polarisierende Vereinfachungen werden gewählt.

**Adjektive:** inkonsistent, wertend, selektiv, reißerisch, spekulativ, unbelegt, pauschal

**Beispiele:**

„Es ist offensichtlich, (etwas soll getan werden)."

„Wie man es kennt, (...)"

## Normativ und pragmatisch

Bezieht sich auf ethische, rechtliche oder gesellschaftlich anerkannte Normen, aber die Argumente sind anfällig für Gegenbeispiele; wissenschaftliche Aussagekraft ist deutlich eingeschränkt. Bewertet Handlungen oder Aussagen nach Nützlichkeit und Machbarkeit. Nicht eindeutige einer Domäne zuzuordnen.

**Adjektive:** werteorientiert, regelgebunden, prinzipientreu, verpflichtend, interpretativ, vorsichtig verallgemeinernd, zweckorientiert, lösungsfokussiert, praktikabel/handlungsorientiert, effizient, faktenbasiert, objektiv

**Beispiele:**

„Derzeit haben wir nicht die Kapazitäten. (Wir wollen das lösen.)"

„Die Infektionszahlen sind gefallen. Die Maßnahmen waren also effektiv."

„Die vorherige Aussage stimmt so nicht. Das wird so nicht funktionieren. (Wir sollen es nicht so umsetzen.)"

„(Es soll online angeboten werden,) damit alle Interessierten teilnehmen können."

„Wenn wir das Problem jetzt ignorieren, wird die Lage sich verschlechtern."

„Arbeit im HomeOffice verringert die Ansteckungsgefahr und auch der Arbeitsweg fällt weg."

„Derzeit ist der Amt in (Bezirk) zu 90% ausgelastet."

„Angenommen eine Person in der Bahn ist infiziert. Dann kann sich das Virus in dem Raum sehr schnell verbreiten. (Deswegen sollen wir das Problem angehen.)"

„Wie mein Vorredner schon gesagt hat, (…)"

## B Prompt Candidates

| | | |
|---|---|---|
| System prompt | | You are a helpful AI assistant specialized on argument detection. You operate in German. |
| Prompt | 1 | Finde argumentative Textstellen zu dem Thema Corona innerhalb des gegebenen Textabschnittes. Ein Argument besteht aus mindestens einem Satz und enthält eine Stellung und eine oder mehrere Begründungen. Es soll so kurz wie möglich gehalten werden und maximal 3 Sätze betragen. Ein Argument soll aus ganzen Sätzen bestehen. Es soll ein Bezug zu Corona haben. Gebe die Argumente als Liste aus. Gebe die Textstellen genauso aus, wie sie sind, ohne Veränderungen. Falls keine Argumente vorhanden sind, soll die Liste leer sein. Keine weiteren Kommentare oder Erkärungen. Finde die Argumente im folgendem Textabschnitt: |
| | 2 | Finde argumentative Textstellen zu dem Thema Corona innerhalb des gegebenen Textabschnittes. Ein Argument besteht aus mindestens einem Satz und enthält eine Stellungsnahme und eine oder mehrere Begründungen. Es soll so kurz wie möglich gehalten werden und maximal 3 Sätze betragen. Ein Argument soll aus ganzen Sätzen bestehen. Es soll ein Bezug zu Corona haben. Gebe die Argumente als Liste aus. Gebe die Textstellen genauso aus, wie sie sind, ohne Veränderungen. Falls keine Argumente vorhanden sind, soll die Liste leer sein. Keine weiteren Kommentare oder Erkärungen. Finde die Argumente im folgendem Textabschnitt: |
| | 3 | Finde argumentative Textstellen zu dem Thema Corona innerhalb des gegebenen Textabschnittes. Ein Argument besteht aus mindestens einem Satz und enthält eine Stellung und eine oder mehrere Begründungen. Es soll so kurz wie möglich gehalten werden und maximal 3 Sätze betragen. Ein Argument soll aus ganzen Sätzen bestehen. Die Stelle muss einen Bezug zu Corona haben. Das kann die Bereiche von Gesundheitsversorgung, Eindämmerungsmaßnahmen, Soziales, Forschung, Internationales oder Logistik betreffen. Insgesamt soll der Umgang mit der Epidemie-Situation oder Handlungen, die aus Pandemie entstanden sind, begründet werden. Gebe die Argumente als Liste aus. Gebe die Textstellen genauso aus, wie sie sind, ohne Veränderungen. Falls keine Argumente vorhanden sind, soll die Liste leer sein. Keine weiteren Kommentare oder Erkärungen. Finde die Argumente im folgendem Textabschnitt: |
| | 4 | Es sollten argumentativen Stellen in dem gegebenen Text gefunden werden. Ein Argument kann ein oder mehrere Sätze enthalten. Ein Argument besteht aus Begründungen und einer Behauptung, die begründet wird. Eine Behauptung ist eine Aussage, die eine Handlung suggeriert. Z.B. Aufforderung, Vorhaben, Standpunkt, Empfehlung/Einschätzung oder Schlussfolgerung. Eine Begründung können Fakten und Bewertungen sein, oder auch z. B. bestehende Probleme, Abwertung anderer Vorschläge, Absicht, Prognosen, Vorteile oder (unabsichtigte Folgen), Veranschaulischungen, Verweise oder offenkundige Gründe sein. Ein Argument soll so kurz wie möglich gehalten werden und maximal 3 Sätze betragen. Die Stelle muss einen Bezug zu Corona haben. Es soll der Umgang mit der Epidemie-Situation oder Handlungen, die aus Pandemie entstanden sind, begründet werden. Gebe die Textstellen genauso aus, wie sie sind, ohne Veränderungen. Gebe die gefundenen Argumente als eine Liste aus: "Argument 1", "Argument 2", ... Falls keine Argumente vorhanden sind, soll de Liste leer sein: [] Finde die Argumente im folgendem Textabschnitt: |
| | 5 | Find argumentative passages within given text. Arguments include a claim and at least one justification. Full reasoning should be included in one argumentative passage. However, the argument should be as short as possible and be 3 sentences max. Do not change any text within the arguments and return found passages as direct quotes. Return all arguments in a list of shape: "Argument 1", "Argument 2", ... If no arguments are found, return an empty list: [ ] Don't provide any comments or explanations. Find argumentative passages in the following text: |
| | 6 | Es sollten argumentativen Stellen, die eine Behauptung mit Begründungen enthalten, in dem gegebenen Text gefunden werden. Ein Argument soll so kurz wie möglich gehalten werden, maximal 3 Sätze. Jede Stelle muss einen Bezug zu Corona haben. Gebe die Textstellen genauso aus, wie sie sind, ohne Veränderungen. Keine sonstigen Kommentare. Falls keine Argumente vorhanden sind, soll de Liste leer sein. Finde die Argumente im folgendem Textabschnitt: |
| | 7 | Find argumentative passages within given text. Arguments include a claim and at least one justification. Full reasoning should be included in one argumentative passage. However, the argument should be as short as possible and be 3 sentences max. Do not change any text within the arguments and return found passages as direct quotes. If no arguments are found, return an empty list. Don't provide any comments or explanations. Find argumentative passages in the following text: |
| | 8 | Find argumentative passages within given text. Arguments include a claim, e.g. a call to action, and at least one justification. Justification can be facts or assessment, for example. Full reasoning should be included in one argumentative passage. However, the argument should be as short as possible and be 3 sentences max. The argumentative passages should be relevant to the Corona discourse and about dealing with the pandemic. Do not change any text within the arguments and return found passages as direct quotes. If no arguments are found, return an empty list. Don't provide any comments or explanations. Find argumentative passages in the following text: |