Misleading through Inconsistency: A Benchmark for Political Inconsistencies Detection

Nursulu Sagimbayeva[1], **Ruveyda Betül Bahçeci**[2], **Ingmar Weber**[1]

[1]Saarland Informatics Campus, Saarland University
[2]Saarland University
nusa00001@uni-saarland.de, ruba00002@uni-saarland.de, iweber@cs.uni-saarland.de

Abstract

Inconsistent political statements represent a form of misinformation. They erode public trust and pose challenges to accountability, when left unnoticed. Detecting inconsistencies automatically could support journalists in asking clarification questions, thereby helping to keep politicians accountable. We propose the Inconsistency detection task and develop a scale of inconsistency types to prompt NLP-research in this direction. To provide a resource for detecting inconsistencies in a political domain, we present a dataset of 698 human-annotated pairs of political statements with explanations of the annotators' reasoning for 237 samples. The statements mainly come from voting assistant platforms such as Wahl-O-Mat in Germany and Smartvote in Switzerland, reflecting real-world political issues. We benchmark Large Language Models (LLMs) on our dataset and show that in general, they are as good as humans at detecting inconsistencies, and might be even better than individual humans at predicting the crowd-annotated ground-truth. However, when it comes to identifying fine-grained inconsistency types, none of the model have reached the upper bound of performance (due to natural labeling variation), thus leaving room for improvement. We make our dataset and code publicly available.[1]

1 Introduction

Once elected, politicians have an unspoken duty to fulfill their campaign promises. While consistency between stated beliefs and legislative actions is associated with credibility and ideological commitment, inconsistencies can undermine public trust and support (Friedman and Kampf 2020).

Figure 1 presents two recent examples of inconsistent statements in German politics. One instance is the Green Party, which abandoned its long-standing opposition to arms exports following Russia's 2022 invasion of Ukraine. Likewise, the AfD, initially opposed to agricultural subsidies, later called for "doubling the diesel refund" in response to large-scale farmer protests, effectively endorsing a form of subsidy. Both cases received significant media attention, highlighting the impact of perceived inconsistencies in public discourse.

[1]https://github.com/nursaltyn/Inconsistency_Detection_Benchmark

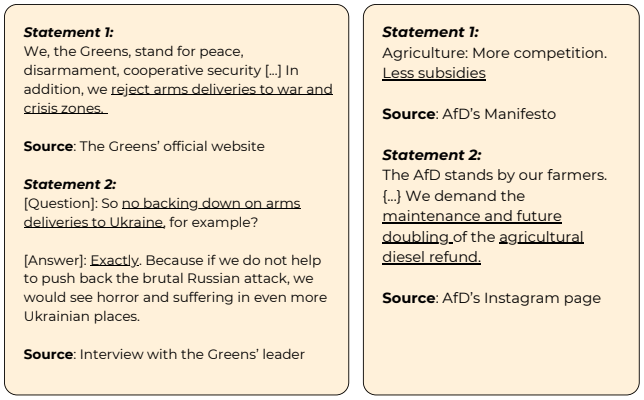


Figure 1: Example of inconsistencies from the Green party (left-wing) and the AfD party (far-right), Germany. Underlined are spans that explain the inconsistencies. Original statements can be found in Appendix A.

Challenging politicians about such inconsistencies is a common tactic for journalists during press conferences to expose contradictions and prompt clarifications. However, with the massive amount of digital content produced by parties and their members, it becomes more challenging to detect inconsistencies manually. Moreover, parties use different platforms to communicate with their electorate, and can alter their messages based on the platform: for example, they can express more aggressive views on TikTok and more moderate on Facebook (McLaughlin et al. 2024). These challenges highlight the need for automated systems to detect inconsistencies.

Inconsistency Detection in political statements extends Natural Language Inference (NLI) task, also known as Recognizing Textual Entailment (RTE). (MacCartney 2009; Dagan and Glickman 2004). While NLI classifies the relationship between two texts as Entailment, Unrelated, or Contradiction, political inconsistency detection involves nuances not fully captured by traditional NLI taxonomies. For example, consider the following statements:

A: Family is the foundation of society and should enjoy special protection and value.

B: Parents should bear the costs of looking after their

children in daycare centers.

At a first glance, *A* and *B* might look unrelated, since there is no direct contradiction between them. However, ideologically they go in opposite directions with respect to family support. One could expect that, everything else being equal, a consistent politician who claims that "family should enjoy special protection and value" will also support free daycare instead of insisting that parents bear the costs. We compare Inconsistency detection to other NLP tasks in Section 3.

In recent years, Large Language Models (LLMs) have shown remarkable capabilities across various NLP tasks (Wei et al. 2022; Zhao et al. 2023), suggesting the potential for deploying Inconsistency detection in real-world scenarios. However, to our knowledge, there are no resources focused on inconsistencies in a political domain. **To address this gap, we introduce a dataset** comprising 698 human-annotated political statements and 334 human explanations for 237 samples (Section 5). The inter-annotator agreement measured by Krippendorff's alpha is **0.528** for the 5-class setting (ordinal scale) and **0.507** for the 3-class setting (nominal scale), which reflects the inherent subjectivity of the task. We evaluate several LLMs on our benchmark and estimate an upper bound of performance for the Inconsistency detection task due to the inherent labeling noise (Section 6).

It is important to emphasize that changing views is not inherently bad and is often unavoidable. Politicians constantly face pressure to accommodate different audiences and react to changes in the environment, which can lead to inconsistencies between their messaging and policies (Friedman and Kampf 2020). In their worst, however, inconsistencies might be a sign of populism (Frisell 2006; Alesina 1988), or even deliberate deception and manipulation. Therefore, Inconsistency detection systems have the potential to promote transparency and accountability.

2 Task formulation

We define our task as follows: given a pair of statements *A* and *B*, detect the relationship between them as one of the labels: *{Unrelated*, *Consistent*, *Inconsistent}*. If the class is Inconsistent, detect a subtype of inconsistency: *{Surface contradiction*, *Factual inconsistency*, *Indirect (Value) inconsistency}*[2]. Under our settings:

- Statements *A* and *B* can be of arbitrary length (from one-liner social media posts to several-page manifestos).
- Statements *A* and *B* might be a part of the bigger document *D*, in which case self-inconsistency detection is performed for *D*.
- To simplify the task, we assume that *A* and *B* were said by the same actor on the same day. This is important because, in reality, time and context significantly influence perceptions of inconsistency. For example, promises made before unexpected crises, such as a pandemic, may become unfulfillable due to changing circumstances, and voters may not view this as inconsistent.

[2]We also considered adding Stereotypes/Expectation violation category, but refrained from it. See details in Appendix B

Our task formulation is versatile and can be applied to various use cases, including detecting inconsistencies:

- Between a party's program or manifesto and its actual policies;
- Across different platforms (e.g., Instagram vs. TikTok);
- Among different party branches (e.g., federal state A vs. federal state B);
- Within a party, among its members.

Types of inconsistency

It's crucial to clarify our definition of "inconsistency" given that numerous casual and philosophical interpretations exist. Inconsistency is not the same as **contradiction**, although contradictions can be viewed as a strong form of inconsistency, where both statements cannot be True at the same time (Dowden 2021). In strict philosophical definitions, statements *A* and *B* are inconsistent with each other if both *cannot* be True, and the truth of one entails the falsity of the other (Wolfram 1989). Our definition of inconsistency, however, allows for situations where the truth of one statement does *not* necessarily entail the falsity of the other. It also permits that both statements could be true or false yet still be inconsistent. To capture the nuances of inconsistency, we conceptualize it as a spectrum with multiple levels (see Table 1). We attach a visual version of our scale in Appendix C.

3 Relation to other work

Contradiction detection is a closely related area of research that focuses on detecting contradictions between a pair of texts (de Marneffe, Rafferty, and Manning 2008). While previous studies have explored Contradiction detection in various domains, such as Finance (Deußer et al. 2023), Medicine (Makhervaks, Gillis, and Radinsky 2023), contracts (Koreeda and Manning 2021), or long documents (Yin, Radev, and Xiong 2021; Li, Raheja, and Kumar 2024), we are not aware of existing NLP research and resources on detecting inconsistencies in political domain. Our work aims to fill this gap by introducing a benchmark dataset consisting of real-world-grounded political statements. In Appendix D, we additionally overview contradiction types used in other papers and map them to classification proposed by us.

Inconsistency Detection Vs Fact-Checking

Fact-checking evaluates the accuracy of statements made by politicians, journalists, and other public figures (Graves and Amazeen 2019). Only claims containing a **purported fact** can be fact-checked (Das et al. 2023). For instance, an opinion such as "Green is the best color" can not be verified since its truthfulness will vary from person to person. Note that in our task, two statements that are both False can either be consistent or inconsistent, depending on their relationship to each other, but independent of an external truth value.

Inconsistency Detection Vs Stance Detection

Stance detection involves identifying an actor's position or attitude toward a specific target topic. Usual labels for attitude are *{Favor, Against, None}* (ALDayel and Magdy

Type	Description	Example	Explanation
Surface contradiction	The strongest degree of inconsistency. No external or specialized knowledge is required to detect it. Understanding logical form and/or language in A and B alone is enough. If A is True, B must be False, and vice versa	Ex. 1) **A:** All kikis are bobable. **B:** This kiki is not bobable. Ex. 2) **A:** We *support* the yellow party. **B:** We *never want to collaborate* with the yellow party.	Ex. 1) We don't have to know what "kiki" and "bobable" mean in the real world. The logical form is: A: All X are Y, B: This Y is not X, and these are mutually exclusive. Ex. 2) Here, we don't have to know who the speaker is, and whether the yellow party exists. From the linguistic formulation only, we can judge that the statements are contradictory.
Factual inconsistency	Having external knowledge about the world *beyond* what is said in A and B is required. This can include laws of physics, economics principles, etc., as well as real-world events and evidence. If A is True, it directly challenges the Truth of B, but does not necessarily make B impossible, and vice versa.	**A:** We will provide extensive social benefits. **B:** We will lower all taxes.	Based on empirical evidence, increasing social benefits usually requires higher taxes. This understanding is necessary to detect inconsistency. At the same time, A and B are not mutually exclusive - the government might also take on more debt or use other ways to increase social benefits. However, based only on the information we have and empirical evidence, we can assume that A and B are Factually inconsistent.
Indirect inconsistency (Value inconsistency)	If A is True, it *doesn't* directly challenge the Truth of B, and vice versa. However, A and B go in *opposite* directions with respect to some value/ideology.	**A:** We voted in favor of increasing data privacy regulations. **B:** We are working on introducing very precise targeted advertising.	In B, one should know that targeted advertising usually relies on extensive user data for higher precision, which inherently conflicts with data privacy. Conversely, A supports data privacy. Thus, it is an Indirect inconsistency.

Table 1: Typology of Inconsistencies

2021). The targets of Stance detection can be either preselected (e.g., given target "Climate change", detect the actor's stance on it) or open-ended, where targets are generated dynamically (Li, Garg, and Caragea 2023). Since Stance detection focuses on speakers' attitudes, it is not very well suited for detecting "neutral" inconsistencies between stated facts such as "combustion engines are contributing to global warming" and "global warming is solely caused by sun spots".

Moreover, targets in contradictions might be highly specific and hard to define, for example:

A: Political parties in Germany shouldn't influence the country's cultural life.

B: Theater ABC in Berlin staged a play financed by party A.

Supposedly, the target here is "Political parties in Germany influencing cultural life", which is hard to account for in advance if we use preset targets. On the other hand, generating such targets dynamically would create a huge universe of possible targets.

Inconsistency Detection Vs Inconsistency Detection in Summarization

The objective of Inconsistency Detection in Summarization (IDS) is to, given the original text and its summary, detect whether the two are consistent or not (Laban et al. 2022; Fabbri et al. 2022; Goyal and Durrett 2020). Some works go beyond binary setting, extending the task to predict the types of factual errors (Chan, Zeng, and Ji 2023; Pagnoni, Balachandran, and Tsvetkov 2021). IDS setting might cover a limited number of contexts in our task: for example, when politicians are directly quoting the manifesto or other sources, one could treat their statement as a summary, and the source they are quoting as the document being summarized.

Still, in our setting, texts A and B are *not* summaries of each other, but rather independent statements. This means that text B can contain novel information not present in A, and vice versa, while still being consistent. However, in the IDS, if a summary contains novel and accurate information not found in the original document, it is labeled as inconsistent (Laban et al. 2022).

4 Sample generation

To create the dataset, we used several approaches: manual and human-LLM collaboration (see Figure 2). We use different pipelines to create samples for different classes, and focus on Inconsistent samples specifically, since we assume that Unrelated and Consistent classes will achieve higher consensus.

Manual pipeline Our initial approach was to find examples of inconsistencies online. We searched on Google formulations such as: "[party_name] contradicts itself" or

"[party_name] being inconsistent" in German. Instead of party_name we inserted names of German parties. Then we read through the articles we found, and either: a) found the original sources of contradictions (e.g. YouTube speech excerpt, Facebook post, etc.) and/or b) summarized the main contradictions in the form "Text 1: A. Text 2: B". Samples obtained via manual pipeline are a minority in the dataset, as collecting them was highly time-consuming, and inconsistencies were relatively rare to find.

Re-using existing datasets Wahl-O-Mat is a German tool for civic education[3] that presents users with socio-political statements, such as "Businesses should be allowed to extend their Sunday opening hours". Users can express their stance as "Agree," "Disagree," or "Neutral." Political parties also declare their positions on these statements. The tool then matches users with the closest political parties. We utilized Qual-O-Mat dataset[4] which contains statements and party's views over various states of Germany from 2002 to 2025 collected from the Wahl-O-Mat application. A notable portion of the statements (2220) also contains comments made by the parties.

We also used X-stance, a stance detection dataset composed of 67,000 comments by Swiss election candidates, addressing over 150 political issues (Vamvas and Sennrich 2020). This data is sourced from the Swiss voting advice application Smartvote[5]. In Figure 15 in the Appendix, we provide a snapshot of both datasets. We also take inspiration from the Perspectrum dataset (Chen et al. 2019), although we do not directly use it in our sample generation pipeline.

Sample generation for each class For Unrelated class, we randomly sample N pairs of statements from Wahl-O-Mat dataset. For Consistent samples, we randomly select N statements and pair each with a comment from the party that supports it[6]. To maintain a standardized sample format, we limit our search to statements where the party's comments are no longer than 160 characters. To generate Surface contradictions, we randomly sample N Wahl-O-Mat statements and match them with N comments from the parties which expressed themselves against the statement. For Factual and Indirect inconsistency, we use several approaches:

- Use LLMs to a) group statements from both datasets into topics (e.g., "Education", "Fiscal Policy") and detect inconsistent texts within each topic; or b) review all presented statements and identify inconsistencies (used for Indirect type; see the prompt in Appendix E);
- Provide LLMs with examples of inconsistent statements to a) generate new inconsistent samples from scratch; b) generate text B in response to input text A such that A and B are inconsistent with each other (used for both Factual and Indirect types; see the prompt in Appendix E);
- Manually create samples by either combining dataset samples or crafting new ones.

We combined manual approach and LLMs for translating, paraphrasing, and filtering samples. Specific names, geographic locations, and party names were removed to mitigate implicit bias. While one might argue that producing statements synthetically from scratch could introduce some leakage into the dataset, most samples were produced by re-sampling other datasets. Moreover, we manually post-edited all samples to ensure broad topic diversity relevant to Germany and Switzerland's political issues.

While we aimed to produce a balanced number of statements that could potentially be labeled as specific classes, the final labels were assigned as a result of crowdsourced annotation (see Section 5). More details on sample generation can be found in Appendix F.

5 Dataset

Our dataset consists of 698 annotated samples, of which 237 samples have at least one explanation for the chosen label. Table 2 reflects a more detailed breakdown of samples by classes. We ensured that all samples from the main dataset were annotated by at least 5 humans.

Inconsistency Type	Count
Surface contradiction	183
Factual inconsistency	122
Indirect inconsistency	179
Consistent	96
Unrelated	118

Table 2: Distribution of final annotations.

Since the perception of inconsistency in politics varies from person to person and might depend on their political views, we expect the Inconsistency detection task to be quite subjective. This reflects in the inter-annotator agreement scores: Krippendorff's alpha $= 0.528$ for 5 classes (ordinal metric), and 0.507 for 3 classes (if we treat all inconsistency types as one Inconsistent class; nominal metric); more on the metric choice in Appendix G. A small subset of samples had more than 5 annotations; in this case, we randomly sampled 5 out of N samples to calculate the agreement. While the score is generally not high, it is common in other subjective tasks, such as judging toxicity of online discussions, emotions by facial expressions, or detecting mental manipulation (Wong, Paritosh, and Aroyo 2021; Wang et al. 2024).

To arrive at the ground truth, we take majority labels for each sample. In case of a tie, which happened in around 16% of the samples, we randomly select one of the top-candidates with equal counts. Both the majority label and the individual labels are part of our dataset.

[3] https://www.bpb.de/themen/wahl-o-mat/

[4] https://github.com/gockelhahn/qual-o-mat-data, shared license-free

[5] https://www.smartvote.ch/de/home, data shared under CC BY-NC 4.0 license

[6] Two authors of the paper manually checked every pair of samples. If we noticed they were correlated (e.g., they could potentially be classified as Consistent/Inconsistent), we re-sampled the statements.

Figure 2: Schematic description of the data annotation pipelines.

Annotation

We recruited crowd-source workers through the Proli c platform[7] and published surveys on Qualtrics[8]. Participants were required to be uent in English and possess at least an undergraduate degree. A basic understanding of politics and economics was recommended. First, we conducted several short rounds of surveys (7 to 10 questions). Since the task is non-trivial, we provided annotators practice sessions to calibrate the understanding of the task. Those who successfully passed comprehension checks were invited to a long study. In total we had 12 annotators evaluate 350 samples each, with one annotator's responses not included in the nal dataset due to suspected inattentiveness. Participants were primarily educated in Social Sciences and Humanities and represented diverse political backgrounds and age groups. Compensation for short studies was at least £9.00 per hour, with the hourly rate increased to £12.21/hour for the long study in accordance with the National Living Wage in the UK for age 21 or over[9]. More details on annotators' demographics and annotation guidelines are available in the Appendix G. We publish the self-reported political leaning together with the annotations to make it possible to evaluate a potential political bias in the annotations.

To prevent the usage of LLMs, we used Proli c's Captcha veri cation feature in the beginning of the surveys. During the long studies, we sometimes asked annotators to brie y explain their reasoning to the question given before, without the option to look back at the previous question. These open-ended questions randomly appeared from 10 to 15% of the time. Additionally, we disabled copy-paste options in the free-text eld.

Repeated samples

Due to an error on our side, 190 samples were accidentally annotated repeatedly several times by the same participants. However, this allowed us to analyze consistency within the same annotator (see Figure 3). Most switches occurred within the ”Inconsistent” class, with the most common case being a shift from Factual to Indirect Inconsistency. We only analyze the rst two answers for each repeatedly annotated sample, since when labeling the same question for the third time, participants most likely have memorized their previous answers. Only the rst answer per sample is included in the dataset.

6 Model evaluation

We evaluated four off-the-shelf models using the same instructions given to annotators, excluding visuals (see prompt in Appendix H). The following models were considered:gpt-4-turbo-2024-04-09 , gpt-3.5-turbo-instruct (further referred to as ChatGPT-4 turbo and ChatGPT-3.5 turbo), LLaMA3.3 70B Instruct, and Llama3 8B Instruct. To obtain predictions for each model, we execute the same prompt ve times to address prediction instability when identical prompts yield

[7] https://www.proli c.com/

[8] https://www.qualtrics.com/

[9] https://www.acas.org.uk/national-minimum-wage-entitlement

Figure 3: Annotation switches within the same participant during repeated trials.

different labels.[10] We then select the majority label and resolve ties by randomly choosing among the top candidate labels. The costs are discussed in Appendix I.

We first evaluate the performance of individual humans in predicting the majority label. For each sample with N human annotations, we used N-1 annotations to determine a majority label, treating the remaining annotation as an individual human's prediction. We repeat this procedure N times, obtaining N pairs of "ground truth" and prediction for each sample. In cases of ties, we randomly selected one of the labels with the highest equal counts as the majority. Overall, this resulted in N×M ground truth-prediction pairs, where N is at least 5 and M is 698, the total number of labeled samples. To evaluate model predictions of the majority label, we use the same setting, comparing model output with N ground-truth labels per sample. We provide an illustration of the evaluation process in Appendix J.

Due to the inherent labeling variation due to subjectivity, achieving 100% accuracy on this task is impossible and would indicate overfitting. To obtain a more realistic upper bound, we simulated a new set of annotators using bootstrapping with replacement. This allowed us to estimate how accurately annotations obtained on a different day could theoretically predict the ground truth established by our current human annotators. For each sample with N annotations, we obtain N tuples of size (N-1). Then for each tuple, we generate 10 boostrapped versions by resampling it with replacement to simulate new annotations. We then determine the majority label for each of the (N-1) x 10 tuples. In this process, we treat one held-out annotation as the ground truth and the 10 bootstrapped majority labels as predictions. Repeating this for each sample, we obtain N x 10 x M tuples, where N is at least 5 and M is 698.

[10] An alternative way to obtain stable prediction would be to set the temperature to 0. We used default temperature and top-P settings.

7 Results

We analyze both 5-class and 3-class settings, where the latter treats all inconsistency types as a single "Inconsistent" class.

In the 3-class setting, both humans and most models seem to approach the upper bound in predicting the general "Inconsistent" class (see Figure 4b). However, in the 5-class setting, neither human annotators nor models reached this upper bound, indicating room for improvement (Figure 4a). Interestingly, performance for most models and humans was lowest for Factual and Indirect inconsistency types, which were also the categories where annotators most frequently changed their minds during repeated trials (Figure 3). Based on F1-score, ChatGPT-4 turbo and LLaMA 70B showed the best overall performance in predicting the majority label, sometimes even outperforming individual human annotators.

Model	Unrel.	Consist.	Inconsist.
ChatGPT-4 turbo	0.548	0.662	0.619
LLaMA 70B	0.525	0.707	0.633
Humans	0.503	0.637	0.617
Bootstrap humans	0.727	0.798	0.786

Table 3: Performance for 3 classes by MCC

Model	Indirect	Factual	Surface
ChatGPT-4 turbo	0.174	0.183	0.328
LLaMA 70B	0.278	0.215	0.388
Humans	0.248	0.256	0.418
Bootstrap humans	0.591	0.573	0.675

Table 4: Performance for Inconsistent class by MCC

We compare the top models using the Matthews Correlation Coefficient (MCC), which has been shown to account for class imbalance more effectively than F1-score (Chicco and Jurman 2020). MCC ranges from -1 to 1, where 1 indicates perfect prediction and -1 indicates an inverse prediction. Tables 3 and 4 feature the results. LLaMA 70B shows a slight advantage over both ChatGPT-4 turbo and humans in predicting Consistent and general Inconsistent classes, while ChatGPT-4 turbo slightly leads in predicting the Unrelated class. LLaMA 70B also excels in Indirect inconsistency predictions, though it remains well below the upper bound. Individual humans remained best in predicting Factual inconsistency and Surface contradiction types.

To further dissect model performance, we analyze Precision vs. Recall scores (Figure 5). We observe that for humans, the difference between precision and recall is relatively modest across all classes. In contrast, for models the

(a) 5 classes

(b) 3 classes

Figure 4: F1-score by model and class.

gap between precision and recall can be stark, especially in the 5-class setting. This indicates that each model has a bias toward certain classes, such as Indirect Inconsistency for LLaMA 70B or Factual Inconsistency for ChatGPT 3.5-turbo - Figure 6 shows high recall while notably lower precision for these classes. The underlying reasons for a certain class preference are not clear; further exploration of models' biases in perceiving inconsistency in politics is a potential future research direction (Section 8).

Figure 5: Precision and Recall distribution for 3 classes.

Figure 6: Precision and Recall distribution for 5 classes.

8 Discussion and future directions

This work and the corresponding dataset release lay the foundation for developing a system for detecting political inconsistencies in-the-wild. One can imagine a pipeline where parties' statements are routinely collected across different platforms (social media, parliament speeches, etc.) and checked for inconsistency with previous statements. To build such a pipeline one has to address challenges related to not knowing which pairs of statement to check for consistency, in particular when dealing with documents of arbitrary length. Promising approaches include pre- ltering by topic to reduce the search space of potential inconsistencies. Absent of a ltering strategy, one would have to naively compare all pairs of statements, leading to scability challenges.

With such a system in place, a large number of empirical studies in computational social science would become feasible, such as comparisons of different parties, countries, and historic times.

9 Limitations

There are several limitations we recognize in our work.

First of all, we acknowledge that perception of inconsistency might be subjective and depend on factors such as personal knowledge, education, political preferences, and even factors such as concentration and attentiveness while judging the statements. Because it is hard to strictly de ne degrees of inconsistency, we believe aligning models' understanding of inconsistency with human understanding is so important; humans as end users of the system should de ne what they care about and would nd useful.

Given that our annotators were predominantly from Europe, it might have in uenced the produced labels. Moreover, quite a large amount of samples to annotate (350 per annotator) could have in uenced their attentiveness. We tried to address it by breaking down the surveys in two parts and randomly asking to explain the annotators' reasoning for the previous question, but the effects might still nd place.

Second of all, we recognize geographical limitations of our dataset: most of the samples were obtained from the context of German and Swiss politics, and the issues discussed in these regions might not be as relevant in other countries. Thirdly, our samples are in English, which might mean that current models would struggle with generalizing to other languages. We see potential improvement in these areas by collecting more diverse samples from countries with different political and economical contexts, and experimenting with generalization of the models to new examples of political inconsistencies.

Another limitation is that our task takes pairs of texts as an input, whereas inconsistency in politics is often also expressed by actions, such as voting for certain law projects. However, if such actions are documented in a form of text, we hope this doesn't pose a major limitation to performing Inconsistency detection.

Lastly, the samples we present are short one-sentence statements. We simplify the setting on purpose to make sure our annotators don't get lost in long documents and overlook potential inconsistencies. In the future, however, we aim to collect longer and noisier text samples and compare model performance.

Ethical Statement

While focusing on bene cial use cases, we recognize that a misuse of Inconsistency detection systems is possible. For example, parties might use such tools to try and harass their opponents. Moreover, when deploying such systems in real world, it would be important to make sure it treats all parties equally, without bias against or in support of any political views.

A broader question that remains is how important consistency in politics is to voters. Recent studies suggest that people tend to overlook political inconsistency as long as the current policy matches their preferences (Croco 2016) and the effects of inconsistency are limited to certain circumstances (Karande, Case, and Mady 2008). Thus, candidates who have been inconsistent might be better off explaining the reasons for their change in position.

Moreover, some studies suggest that politicians who are more averse to lying have lower reelection rates, indicating that honesty might not pay off in politics (Janezic and Gallego 2020). Still, we believe that due to information overload and scarce resources to monitor political parties' online presence attentively, many inconsistencies go unnoticed, and at least detecting them would be bene cial for both practical and research purposes.

Regarding the content of the dataset, we realize that it contains some samples that some people might nd offensive (for example, statements like "Headscarves for female teachers in public schools must be banned."). The statements presented in the dataset do not re ect personal views of the authors. The majority of the statements comes from the of cial voting assisting tools, such as Wahlomat and Smartvote.

Regarding the annotation process, we received feedback that the task is rather intense, and therefore we broke down the main annotation into two parts. The annotators were given a week to annotate each part in their own tempo (not all questions at once), and were strongly encouraged to take a break when they start to get tired.

Acknowledgements

Nursulu Sagimbayeva and Ingmar Weber are supported by funding from the Alexander von Humboldt Foundation and its founder, the Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung). We would like to thank our colleagues from the Interdisciplinary Institute for Societal Computing for their thoughtful feedback and suggestions.

We used AI tools to assist with tasks such as literature search (Elicit.ai[11]) and minor coding and paraphrasing tasks (ChatGPT, Mistral). Every suggestion made by AI tools was veri ed by the authors.

References

Aharoni, R.; Narayan, S.; Maynez, J.; Herzig, J.; Clark, E.; and Lapata, M. 2023. Multilingual Summarization with Factual Consistency Evaluation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., Findings of the Association for Computational Linguistics: ACL 2023, 3562–3591. Toronto, Canada: Association for Computational Linguistics.

ALDayel, A.; and Magdy, W. 2021. Stance detection on social media: State of the art and trends. Information Processing & Management, 58(4): 102597.

Alesina, A. 1988. Credibility and Policy Convergence in a Two-Party System with Rational Voters. The American Economic Review, 78(4): 796–805.

Chan, H. P.; Zeng, Q.; and Ji, H. 2023. Interpretable Automatic Fine-grained Inconsistency Detection in Text Summarization. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., Findings of the Association for Computational Linguistics: ACL 2023, 6433–6444. Toronto, Canada: Association for Computational Linguistics.

Chen, S.; Khashabi, D.; Yin, W.; Callison-Burch, C.; and Roth, D. 2019. Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. In Conference

[11] https://elicit.com/

of the North American Chapter of the Association for Computational Linguistics (NAACL).

Chicco, D.; and Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1): 6.

Cohen, R.; Hamri, M.; Geva, M.; and Globerson, A. 2023. LM vs LM: Detecting Factual Errors via Cross Examination. In Bouamor, H.; Pino, J.; and Bali, K., eds., Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 12621–12640. Singapore: Association for Computational Linguistics.

Croco, S. E. 2016. The Flipside of Flip-Flopping: Leader Inconsistency, Citizen Preferences, and the War in Iraq. Foreign Policy Analysis, 12(3): 237–257.

Dagan, I.; and Glickman, O. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. Proceedings of the PASCAL Workshop on Textual Entailment and Paraphrasing. This is a workshop paper and might not have a standard journal reference.

Das, A.; Liu, H.; Kovatchev, V.; and Lease, M. 2023. The state of human-centered NLP technology for fact-checking. Information Processing & Management, 60(2): 103219.

de Marneffe, M.-C.; Rafferty, A. N.; and Manning, C. D. 2008. Finding Contradictions in Text. In Moore, J. D.; Teufel, S.; Allan, J.; and Furui, S., eds., Proceedings of ACL-08: HLT, 1039–1047. Columbus, Ohio: Association for Computational Linguistics.

DellaPosta, D.; Shi, Y.; and Macy, M. 2015. Why Do Liberals Drink Lattes? American Journal of Sociology, 120(5): 1473–1511.

Deußer, T.; Pielka, M.; Pucknat, L.; Jacob, B.; Khameneh, T. D.; Nourimand, M.; Kliem, B.; Loitz, R.; Bauckhage, C.; and Sifa, R. 2023. Contradiction Detection in Financial Reports. Proceedings of the Northern Lights Deep Learning Workshop.

Dowden, B. 2021. Recognizing Inconsistency and Contradiction. Accessed: February 4, 2025.

Fabbri, A.; Wu, C.-S.; Liu, W.; and Xiong, C. 2022. QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2587–2601. Seattle, United States: Association for Computational Linguistics.

Friedman, E.; and Kampf, Z. 2020. `To thine own self be true': The perceived meanings and functions of political consistency. Language in Society, 49(1): 89–113.

Frisell, L. 2006. Populism. Microeconomic Theory eJournal.

Goyal, T.; and Durrett, G. 2020. Evaluating Factuality in Generation with Dependency-level Entailment. In Cohn, T.; He, Y.; and Liu, Y., eds., Findings of the Association for Computational Linguistics: EMNLP 2020, 3592–3603. Online: Association for Computational Linguistics.

Graves, L.; and Amazeen, M. 2019. Fact-checking as idea and practice in journalism. In Nussbaum, J., ed., Fact-Checking and the Future of Journalism. Oxford University Press.

Huntsman, S.; Robinson, M.; and Huntsman, L. 2024. Prospects for inconsistency detection using large language models and sheaves. arXiv:2401.16713.

Janezic, K. A.; and Gallego, A. 2020. Eliciting preferences for truth-telling in a survey of politicians. Proceedings of the National Academy of Sciences, 117(36): 22002–22008.

Karande, K.; Case, F. M.; and Mady, T. 2008. When does a candidate's inconsistency matter to the voter? International Journal of Advertising, 27: 37 – 65.

Koreeda, Y.; and Manning, C. 2021. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., Findings of the Association for Computational Linguistics: EMNLP 2021, 1907–1919. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Laban, P.; Kryściński, W.; Agarwal, D.; Fabbri, A. R.; Xiong, C.; Joty, S.; and Wu, C.-S. 2023. LLMs as Factual Reasoners: Insights from Existing Benchmarks and Beyond. arXiv:2305.14540.

Laban, P.; Schnabel, T.; Bennett, P. N.; and Hearst, M. A. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. Transactions of the Association for Computational Linguistics, 10: 163–177.

Lattimer, B.; Chen, P. H.; Zhang, X.; and Yang, Y. 2023. Fast and Accurate Factual Inconsistency Detection Over Long Documents. In Bouamor, H.; Pino, J.; and Bali, K., eds., Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 1691–1703. Singapore: Association for Computational Linguistics.

Li, J.; Raheja, V.; and Kumar, D. 2024. ContraDoc: Understanding Self-Contradictions in Documents with Large Language Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 6509–6523. Mexico City, Mexico: Association for Computational Linguistics.

Li, Y.; Garg, K.; and Caragea, C. 2023. A New Direction in Stance Detection: Target-Stance Extraction in the Wild. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 10071–10085. Toronto, Canada: Association for Computational Linguistics.

Lin, G.; and Zhang, Y. 2023. Sparks of Artificial General Recommender (AGR): Experiments with ChatGPT. Algorithms, 16(9): 432.

MacCartney, B. 2009. Natural Language Inference. Ph.D. thesis, Stanford University.

Makhervaks, D.; Gillis, P.; and Radinsky, K. 2023. Clinical Contradiction Detection. In Bouamor, H.; Pino, J.; and Bali,

K., eds., Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 12248–12263. Singapore: Association for Computational Linguistics.

McLaughlin, B.; Cloudy, J.; Hunter, J.; and Potter, B. 2024. Stitch incoming: political engagement and aggression on TikTok. Behaviour and Information Technology.

Pagnoni, A.; Balachandran, V.; and Tsvetkov, Y. 2021. Understanding Factuality in Abstractive Summarization with FRANK: A Benchmark for Factuality Metrics. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4812–4829. Online: Association for Computational Linguistics.

Sepúlveda-Torres, R.; Bonet-Jover, A.; and Saquete, E. 2021. "Here Are the Rules: Ignore All Rules": Automatic Contradiction Detection in Spanish. Applied Sciences, 11(7): 3060.

Sepúlveda-Torres, R.; Bonet-Jover, A.; and Saquete, E. 2023. Detecting Misleading Headlines Through the Automatic Recognition of Contradiction in Spanish. IEEE Access, 11: 72007–72026.

Shan, G.; Zhou, L.; and Zhang, D. 2021. From conflicts and confusion to doubts: Examining review inconsistency for fake review detection. Decision Support Systems, 144: 113513.

Vamvas, J.; and Sennrich, R. 2020. X-Stance: A Multilingual Multi-Target Dataset for Stance Detection. arXiv:2003.08385.

Wang, Y.; Yang, I.; Hassanpour, S.; and Vosoughi, S. 2024. MentalManip: A Dataset For Fine-grained Analysis of Mental Manipulation in Conversations. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 3747–3764. Bangkok, Thailand: Association for Computational Linguistics.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022. Emergent Abilities of Large Language Models. ArXiv, abs/2206.07682.

Wolfram, S. 1989. Philosophical Logic. Routledge.

Wong, K.; Paritosh, P.; and Aroyo, L. 2021. Cross-replication Reliability - An Empirical Approach to Interpreting Inter-rater Reliability. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 7053–7065. Online: Association for Computational Linguistics.

Xu, L.; Su, Z.; Yu, M.; Xu, J.; Choi, J. D.; Zhou, J.; and Liu, F. 2024. Identifying Factual Inconsistencies in Summaries: Grounding LLM Inference via Task Taxonomy. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., Findings of the Association for Computational Linguistics: EMNLP 2024, 14626–14641. Miami, Florida, USA: Association for Computational Linguistics.

Xue, T.; Wang, Z.; Wang, Z.; Han, C.; Yu, P.; and Ji, H. 2023. RCOT: Detecting and Rectifying Factual Inconsistency in Reasoning by Reversing Chain-of-Thought. arXiv:2305.11499.

Yin, W.; Radev, D.; and Xiong, C. 2021. DocNLI: A Large-scale Dataset for Document-level Natural Language Inference. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 4913–4922. Online: Association for Computational Linguistics.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.; and rong Wen, J. 2023. A Survey of Large Language Models. ArXiv, abs/2303.18223.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair pro ling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately re ect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we explain the model evaluation methodology and reasoning behind selected methods in section 6. Model evaluation
   (d) Do you clarify what are possible artifacts in the data used, given population-speci c distributions? Yes, we talk about potential sampling bias in section 9. Limitations
   (e) Did you describe the limitations of your work? Yes, section 9. Limitations
   (f) Did you discuss any potential negative societal impacts of your work? Yes, section 9 Ethical Statement
   (g) Did you discuss any potential misuse of your work? Yes, section 9 Ethical Statement
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of ndings? Yes, section 4. Sample generation (we anonymize names of politicians and parties)
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justi cations for all theoretical results? N/A

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A

(e) Did you address potential biases or limitations in your theoretical framework? N/A

(f) Have you related your theoretical results to the existing literature in social science? N/A

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? N/A

(b) Did you include complete proofs of all theoretical results? N/A

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, we included prompts in Appendices F, H, and we make our dataset and code publically available

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? No

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, Appendix I

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, section 6. Model evaluation

(f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? No

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, without compromising anonymity...

(a) If your work uses existing assets, did you cite the creators? Yes, section 4. Sample generation

(b) Did you mention the license of the assets? Yes, section 4. Sample generation

(c) Did you include any new assets in the supplemental material or as a URL? Yes, we included a link to our dataset and code in the abstract

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? No, we didn't explicitly ask for consent of annotators, however, we are not using the personal data of the annotators, but labels that they assigned to statements. We mentioned in the study description that we are conducting an Inconsistency detection task.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we discussed potential content offensiveness in 9 Ethical Statement, and we did not use personally identifiable data

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? No

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, we addressed the questions it asks in our document in section 5. Dataset, but we did not literally fill out the datasheet

6. Additionally, if you used crowdsourcing or conducted research with human subjects, without compromising anonymity...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, Appendix G. Annotation guidelines

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No, all tasks were designed to be low-risk, involving only non-sensitive political texts from public sources.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, section 5. Dataset. Annotation

(d) Did you discuss how data is stored, shared, and deidentified? No, as we do not store personally identifiable participant data

Appendix

A Original statements from German politics examples (Figure 1)

Statements from the Green party

Text 1: Wir Grüne im Bundestag stehen für Frieden, Abrüstung, kooperative Sicherheit und eine Kultur der militärischen Zurückhaltung sowie eine Stärkung der Parlamentsrechte. Unsere Politik zielt darauf ab, Konflikte gar nicht erst entstehen zu lassen. Wir fordern, die zivile Krisenprävention ins Zentrum deutscher Außenpolitik zu stellen und sich engagiert für internationale Abrüstung und Rüstungskontrolle einzusetzen. Wir unterstützen das Recht jedes Landes auf Selbstverteidigung nach Artikel 51 der VN-Charta. Darüber hinaus lehnen wir Waffenlieferungen in Kriegs- und Krisengebiete ab.

(Source: https://www.gruene-bundestag.de/themen/sicherheitspolitik#:~:text=Dar%C3%BCber%20hinaus%20lehnen%20wir%20Waffenlieferungen%20in%20Kriegs-%20und%20Krisengebiete%20ab)

Note: the original text was edited at the party's website as the paper was being written.

Text 2:

Question: Frage: Also kein Zurückweichen etwa bei Waffenlieferungen an die Ukraine?

Answer: Außenministerin Annalena Baerbock: Genau. Denn wenn wir nicht helfen, den brutalen russischen Angriff zurückzudrängen, würden wir an noch mehr ukrainischen Orten Horror und Leid sehen.

(Source: https://www.auswaertiges-amt.de/de/newsroom/interview-aussenministerin-baerbock-sz/2557862)

Statements from the AfD

Text 1: 13.6 Landwirtschaft: Mehr Wettbewerb. Weniger Subventionen

(Source: https://www.afd.de/wp-content/uploads/2023/05/Programm_AfD_Online_.pdf)

Text 2: Die AfD steht an der Seite unserer Landwirte. [...] Wir fordern: Die Verdopplung der Agrardiesel-Rückerstattung.

(Source: https://www.afd.de/sofortprogramm-landwirtschaft/)

B Inconsistency type not included in the final scale

One more category that we considered adding is Stereotypes violation/Expectation violation, where the inconsistency cannot be properly explained logically and relies purely on one's expectations and experience (Dowden 2021). Consider the following statements:

A: I am a Democrat. B: I don't like latte.

There exists a stereotype that US Democrats tend to drink latte more often than Republicans (DellaPosta, Shi, and Macy 2015). However, it is hard to explain logically or ideologically why not liking latte as a Democrat is inconsistent. We refrained from including such inconsistencies into our classification scale due to a high subjectivity of this class and lack of reliable sources of politics-related stereotypes. Hence, such cases should be labeled as "Unrelated" according to our taxonomy.

C Scale visualization

See Figure 7.

D Overview of contradiction types

See Figure 5 for an overview of contradiction types in other literature and its relation to our taxonomy. Works marked with an asterisk * name their category differently, but based on their meaning, we categorize them as Factual inconsistency type.

There are some types of contradictions used in other papers that we don't include in our classification.

Perspective / View / Opinion and Emotion / Mood / Feeling contradiction types used in (Li, Raheja, and Kumar 2024) do not strictly fall into one of the inconsistency categories defined by us; they are rather orthogonal to our scale. For example, sentences:

A) We like the Blue party.

B) We hate the Blue party.

could be labeled as Surface contradiction, whereas sentences:

A)We are keen on forming a coalition with the Blue party.

B) We accuse John Smith of fraud in the election (given that John Smith is a leader of the Blue party).

also signal shift in emotion/perspective, but would probably fall under Factual inconsistency category since noticing this inconsistency requires world knowledge about party leaders.

Another type that we do not include is Structure contradiction used in (de Marneffe, Rafferty, and Manning 2008; Sepúlveda-Torres, Bonet-Jover, and Saquete 2023; Sepúlveda-Torres, Bonet-Jover, and Saquete 2021). Structure contradictions are such that the structure of one of the sentences is not compatible with the other. This is achieved, for example, by interchanging named entities in the sentences:

A) On January 26, 2014, Google acquired DeepMind.

B) On January 26, 2014, DeepMind acquired Google.

However, structure contradiction does not strictly fit one of the contradiction types that we adopt. For example, sentences:

A) Kiki wugged Boba.

B) Boba wugged Kiki.

would fall under Surface contradiction, whereas sentences:

A) Germany gave a large loan to Greece.

B) Greece gave a large loan to Germany.

would fall under the Factual inconsistency category, since noticing inconsistency requires knowledge about the economy and relations between these countries.

Moreover, sometimes change in structure doesn't lead to a contradiction:

A) Merkel gave a present to Scholz.

B) Scholz gave a present to Merkel.

Here, without having additional context, these two could happen simultaneously and are not incompatible.

We also did not include contradiction typology from domains that work with features beyond text, for example, rating-sentiment inconsistency (Shan, Zhou, and Zhang 2021).

Figure 7: Inconsistency scale visualization

Type (our scale)	Other names	Example	Literature
Surface contradiction	Logical	1. All houses are green 2. Sally's house is red.	(Lin and Zhang 2023)
	Negation	1) Sally donated her kidney. 2) Sally never donated her kidney.	(de Marneffe, Rafferty, and Manning 2008), (Sepúlveda-Torres, Bonet-Jover, and Saquete 2023), (Li, Raheja, and Kumar 2024), (Sepúlveda-Torres, Bonet-Jover, and Saquete 2021)
	Numeric	1) More than 50 civilians tragically died as a result of explosion. 2) The police found 32 confirmed dead so far.	(de Marneffe, Rafferty, and Manning 2008), (Sepúlveda-Torres, Bonet-Jover, and Saquete 2023), (Li, Raheja, and Kumar 2024), (Sepúlveda-Torres, Bonet-Jover, and Saquete 2021), (Deußer et al. 2023)
	Antonyms	1) Capital punishment is a catalyst for more crime. 2) Capital punishment is a deterrent to crime.	(de Marneffe, Rafferty, and Manning 2008), (Sepúlveda-Torres, Bonet-Jover, and Saquete 2023), (Sepúlveda-Torres, Bonet-Jover, and Saquete 2021)
	Lexical	1) The Canadian parliament's Ethics Commission said former immigration minister, Judy Sgro, did nothing wrong and her staff had put her into a conflict of interest. 2) The Canadian parliament's Ethics Commission accuses Judy Sgro.	(de Marneffe, Rafferty, and Manning 2008)
	Content	1) She donated her kidney to a stranger. 2) She donated her kidney to a close friend.	(Li, Raheja, and Kumar 2024)
	Factive (Exaggeration)	1) Isuzu and Volvo agree to create a strategic alliance in heavy duty trucks. 2) Isuzu and Volvo create a strategic alliance in heavy duty trucks.	(Sepúlveda-Torres, Bonet-Jover, and Saquete 2023)
Factual inconsistency	Factual	Abraham Lincoln is my mother.	(de Marneffe, Rafferty, and Manning 2008) *, (Huntsman, Robinson, and Huntsman 2024)*, (Aharoni et al. 2023), (Cohen et al. 2023), (Laban et al. 2023), (Xue et al. 2023), (Lattimer et al. 2023), (Xu et al. 2024), (Makhervaks, Gillis, and Radinsky 2023), (Lin and Zhang 2023)
	World Knowledge	1) Microsoft Israel, one of the first Microsoft branches outside the USA, was founded in 1989. 2) Microsoft was established in 1989.	(de Marneffe, Rafferty, and Manning 2008)
	Semantic	1) On 14th of March, 2020, we increased our capital by offering 5,000 new shares during a seasoned equity offering. 2) During 2020 we did not increase our total amount of equity and thus, it remained unchanged at $10,000,000.	(Deußer et al. 2023)
	Causal	1) I slam the door. 2) After I do that, the door opens.	(Li, Raheja, and Kumar 2024)
	Relation	1) Jane and Tom are a married couple. 2) Jane is Tom's sister.	(Li, Raheja, and Kumar 2024)
Value inconsistency	Violation of expectations	I didn't attend the funeral, but I sent a nice letter saying I approved of it.	(Lin and Zhang 2023)

Table 5: Types of Inconsistency in other literature.

E Prompts for sample generation

Prompt for generating Factually inconsistent statement B for input A See Figure 8.

Prompt for detecting Indirect inconsistencies given a set of political statements See Figure 9.

F Sample generation details

Two authors of the paper manually checked every pair of samples.

When randomly sampling Unrelated statements, we re-sampled them if we noticed they were correlated (e.g., they could potentially be classi ed as Consistent/Inconsistent).

During the random sampling of Unrelated statements, any samples that were correlated (e.g. they could be classi ed as Consistent or Inconsistent) were excluded and replaced through re-sampling.

For producing Factual and Indirect inconsistencies, we used gpt-4-0613 and gpt-4o-2024-11-20| with temperature=1.0, top-P=1.0.

Proportion of data produced by each method

We estimate that around 50 samples were produced manually, by summarizing existing party contradictions or issues highlighted in other sources such as petitions from Change.org. Out of them, a subset of around 30 samples was used for annotation guidelines, prompting LLMs, and practice sessions for annotators. Thus, we did not include this subset in the nal dataset to prevent data leakage.

Around 80 samples were produced synthetically via prompting LLMs with examples of different Inconsistency classes. This low number is a result of stringent ltering since many LLM-produced samples were repetitive or did not seem complex enough for the Factual or Indirect Inconsistency classes.

The rest of the samples were produced by reusing the Wahl-O-Mat and X-stance datasets, and considerable manual labor was also invested in re ning and rephrasing the statements.

G Annotation details

Inter-annotator agreement

To measure agreement across ve classes, we use Krippendorff's alpha with an ordinal metric, acknowledging that class order is important. For instance, misclassifying Surface contradiction as Factual inconsistency should be penalized less than misclassifying it as Consistent. For the three classes—Unrelated, Consistent, and Inconsistent — we applied Krippendorff's alpha with a nominal metric, as these categories are best treated as purely categorical without an inherent order.

Annotator Demographics

Most participants resided in Europe, and some in the UK and the US. The majority of participants had obtained their education in Social Sciences and Humanities. To estimate possible bias in evaluating political statements, we asked participants to indicate their political standing. We obtained the following distribution of self-reported political views: Left-wing - 27%, Rather left-leaning - 18%, Moderate - 27%, Rather right-leaning - 27%. Age groups were distributed the following way: 18-24 - 4 people, 25-34 - 4 people, 35-44 - 2 people, 55-64 - 1 person.

By trying to label some samples ourselves, we estimated that annotating each sample takes around 1 minute on average.

Annotation guidelines

After providing instructions in short studies, we conducted a brief comprehension check consisting of four questions. Only participants who scored at least 2/4 were eligible for further consideration.

To ensure a better understanding of the task before comprehension checks, we also conducted practice sessions, where participants could annotate a set of questions and see the answers and explanations we provided for them. The study was split into two parts, with a total of 350 samples per participant. We recruited 12 annotators for the main study. The answers from one annotator were discarded upon reviewing due to the explanations being too generic and possibly showing inattentiveness. Additionally, one participant chose not to participate in part 2 of the study, and was replaced by another annotator.

We attach screenshots from annotation guidelines in Figures 10, 11, 12, 13.

Figure 8: Prompt for generating Factual inconsistencies).

Figure 9: Prompt for generating Indirect inconsistencies.

Figure 10: Instructions for annotators (Page 1)

Figure 11: Instructions for annotators (Page 3)

Figure 12: Instructions for annotators (Page 4)

Survey interface

See Figure 14.

Samples from Wahl-O-Mat and X-stance

See Figure 15.

H Prompt for model evaluation

”You are a diligent annotator who is performing a crowdsourcing task. You will be given pairs of texts, which are statements made by a political party. Your task is to evaluate whether they are consistent with each other or not. Additionally, we will ask you to briefly explain your reasoning behind choosing the label.

Please familiarize yourself with the evaluation scale. Evaluation scale (In growing order, the explanations will follow) Unrelated - Consistent - Indirect inconsistency - Factual inconsistency - Surface contradiction

Surface contradiction

If A is True, B is False, and vice versa. No external/specialized knowledge is required to detect contradiction, just understanding logical form/language in A and B.

Example 1: a) All kikis are bobable. b) This kiki is not bobable. Explanation: the logical form here is: a) All A are B. b) This A is not B. These are mutually exclusive.

Example 2: a) I love kikis. b) I think kikis are the most terrible thing on Earth. Explanation: knowing language here is enough to see a contradiction.

In both examples, we don’t have to know what ”kiki” and ”bobable” mean in the real world (no external knowledge is needed). The ”surface” is enough to see a contradiction.

Example 3: a) I support the yellow party. b) I’m against the yellow party.

Explanation: To support something and to be against something are the opposites. It is enough to understand the language to detect a contradiction - thus, it is a Surface contradiction.

Example 4: a) We oppose sending weapons to war zones. b) We voted in favor of sending 50 tanks to a country that was attacked.

Explanation: To ”oppose sending weapons” and ”vote in favor of sending 50 tanks” express opposing attitudes. No knowledge beyond A and B is needed to see a contradiction, so it is a Surface contradiction.

Factual inconsistency

if A is True, it challenges the Truth of B. Having external knowledge about the world beyond what is said in A and B is required to see inconsistency. This knowledge can include laws of physics, principles of economics, international relations, etc., as well as real-world events and empirical evidence.

Example 1: a) We will provide extensive social benefits. b) We will minimize all the taxes.

Explanation: based on empirical evidence, increasing social benefits (pensions, subsidies, etc.) usually requires having high taxes. This knowledge is needed to detect inconsistency.

Note that A and B are not mutually exclusive - maybe the government will take debt or use other ways to increase social benefits. However, based only on the information we have + empirical evidence, we can assume that A and B are Factually inconsistent.

Example 2: a) We care about climate change and want to switch to renewable energy. b) We are planning to build 120 coal plants next year. Explanation: Coal is not a renewable energy source, and burning coal contributes to climate change. This knowledge is required to see the inconsistency, and it is not mentioned in A and B. Thus, A and B are Factually inconsistent.

Example 3: a) We don’t support subsidies in agriculture and think they are bad for competition. b) We demand the maintenance and future doubling of the agricultural diesel refund.

Explanation: Text A is against subsidies in agriculture. Text B supports diesel refund, which can be considered as a type of a subsidy. Knowing this is required to see a contradiction, thus, it is a Factual inconsistency.

Indirect inconsistency

If A is True, it doesn’t directly challenge the Truth of B, and vice versa. However, A and B go in opposite directions with respect to some value/ideology (V).

Example 1: a) We support financial aid for unemployed people. b) We are against financial aid for single mothers.

Explanation: In A, the author supports financial aid for people who might be struggling financially, while B is implicitly against supporting them (since single mothers might be struggling financially too). However, A doesn’t challenge the truth of B - they are just going in opposite political directions.

Example 2: a) We voted in favor of increasing data privacy regulations. b) We are working on introducing very precise targeted advertising. Explanation: In text B, one should know that targeted advertising requires extensive user data to be more precise. Thus, text B goes against data privacy, while text A supports it. It is an Indirect inconsistency.

Example 3: a) Our government is committed to leading the disarmament negotiations. b) The defense ministry has announced an increase in the benefits available for volunteers in the army. Explanation: Text A expresses a stance against militarization (leading the disarmament negotiations), while text B is pro-militarization (recruiting more volunteers in the army). Thus, there is an Indirect inconsistency between A and B.

Consistent

If A is True, B is also likely to be True, and vice versa.

Example 1: a) I like classical music. b) I am going to a classical music concert today.

Example 2: a) We believe that the current government is failing to address population decline effectively. b) We believe the government should prioritize the traditional family model.

Example 3: a) We believe that determining asylum eligibility before individuals reach the country would significantly relieve the taxpayers. b) We support stricter border protection measures due to the large-scale family reunification migration.

Unrelated

Consistent or Unrelated?

It might be so that the statements are on the same topic, but still Unrelated.

Example: Education
a) We support standardized school uniforms**. b)** We want to introduce the 13-th grade in schools.

Think about it in the following way: if the truth of A changes (e.g. True -> False), does it **affect** the likelihood of B? (in any direction, i.e. makes it more/less likely)
If **not**, then the statements are **Unrelated**.

Confused between two categories?

If you think the detected contradiction fits more than one category**, choose the category that is more to the right of the spectrum**.

E.g. if you think the inconsistency could be both Factual and Surface, choose Surface.

Important

When you decide whether statements A and B are inconsistent, imagine that the **same politician** said them **on the same day,** and then make a decision.

This is because some statements could be consistent if said with a difference in 2 years, but inconsistent when said on the same day.

Figure 13: Instructions for annotators (Page 5)

The Truth of A doesn't affect the Truth of B, and vice versa.

Example 1: a) All apples are red. b) I like sunny weather.

Example 2: a) We think people should be able to obtain sick leave for mild illnesses for up to seven days. b) We believe that the left-leaning bias in public broadcasting journalism is a significant issue.

Example 3: a) We oppose expanding surveillance measures that use facial recognition technology. b) We believe that the deportation of illegal migrants will help alleviate the housing shortage in our country.

Now you will be given a pair of statements. Evaluate them, and output only the label and the explanation in a format:

Label: your label Explanation: your explanation ********

Input texts: "

I Model running costs

We didn't keep precise track of our model evaluation budget. However, based on the number of tokens, we estimate each complete run over 698 samples with OpenAI API for ChatGPT-4 turbo and ChatGPT-3.5 turbo together costed us around $13. The prices might differ to some extent in reality. Moreover, under our prompt, the model outputs label **and its explanation,** which might have increased the costs. Below we list the prices relevant to the date:

- **ChatGPT4-turbo:** $10.00 / 1M input tokens and $30.00 / 1M output tokens.
- **ChatGPT3.5-turbo instruct:** $1.50 / 1M input tokens and $2.00 / 1M output tokens.

Up-to-date prices can be looked up at OpenAI's official website: https://openai.com/api/pricing/.

For using LLaMA models InferenceAPI, we paid for the Pro account subscription on HuggingFace, which at the moment of writing the paper costed $9. Up-to-date prices can be looked up at HuggingFace's official website: https://huggingface.co/pricing.

J Evaluation visualization (majority class prediction)

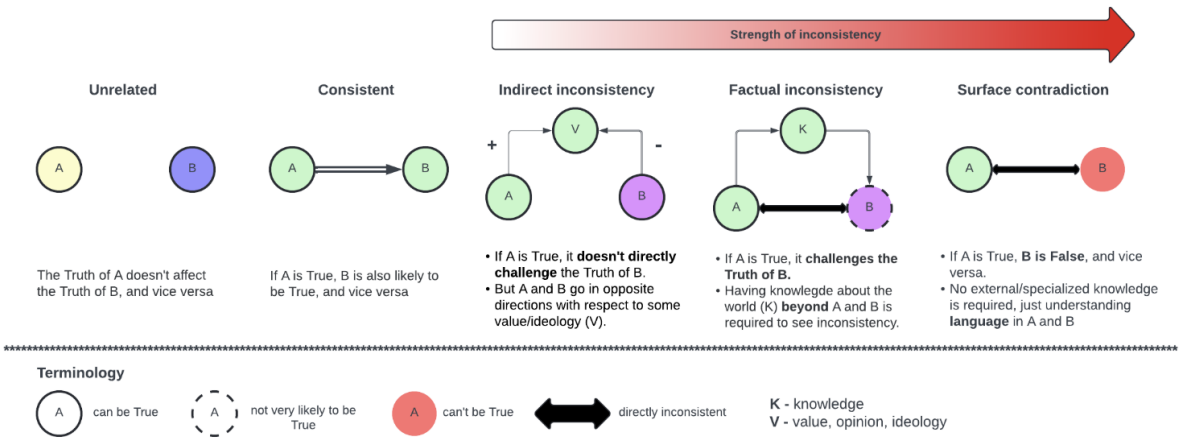
See Figure 16.

Now let's start the practice session (6 questions).

Remember:

- Always imagine the two statements were made by the **same politician** on the **same day.**
- If **undecided** between two categories, choose the one **more to the right** of the spectrum.

Q1.
a) We care about climate change and want to switch to renewable energy.
b) We are planning to build 120 coal plants next year.



Unrelated	Consistent	Indirect inconsistency	Factual inconsistency	Surface contradiction
○	○	○	○	○

→

& Terms

Figure 14: Example from a practice session