

Hybrid Annotation for Propaganda Detection: Integrating LLM Pre-Annotations with Human Intelligence

Ariana Sahitaj^{1,2*} Premtim Sahitaj^{1,2*} Veronika Solopova^{1,2}
Jiaao Li^{1,2} Sebastian Möller^{1,2} Vera Schmitt^{1,2}

¹Quality and Usability Lab, Technische Universität Berlin, Germany

²German Research Center for Artificial Intelligence (DFKI), Berlin, Germany

ariana.sahitaj@campus.tu-berlin.de

Abstract

Propaganda detection on social media remains challenging due to task complexity and limited high-quality labeled data. This paper introduces a novel framework that combines human expertise with Large Language Model (LLM) assistance to improve both annotation consistency and scalability. We propose a hierarchical taxonomy that organizes 14 fine-grained propaganda techniques (Martino et al., 2020) into three broader categories, conduct a human annotation study on the HQP dataset (Maarouf et al., 2023) that reveals low inter-annotator agreement for fine-grained labels, and implement an LLM-assisted pre-annotation pipeline that extracts propagandistic spans, generates concise explanations, and assigns local labels as well as a global label. A secondary human verification study shows significant improvements in both agreement and time-efficiency. Building on this, we fine-tune smaller language models (SLMs) to perform structured annotation. Instead of fine-tuning on human annotations, we train on high-quality LLM-generated data, allowing a large model to produce these annotations and a smaller model to learn to generate them via knowledge distillation. Our work contributes towards the development of scalable and robust propaganda detection systems, supporting the idea of transparent and accountable media ecosystems in line with SDG 16. The code is publicly available at our GitHub repository¹.

Content Warning: This paper contains examples of Russian propaganda, some of which contain misleading, or offensive claims. These are provided for academic analysis and do not reflect the authors' views.

```
"input": "So what if Zelensky is Jewish [...] The fact
does not negate the Nazi elements in Ukraine. I believe
that Hitler also had Jewish blood.",
"output": {
  "propaganda_spans": [
    {
      "span": "The fact does not negate the Nazi
elements in Ukraine",
      "explanation": "[...]",
      "local_label": "doubt"
    },
    {
      "span": "I believe that Hitler also had Jewish
blood",
      "explanation": "[...]",
      "local_label": "reductio_ad_hitlerum"
    }
  ],
  "global_label": "reductio_ad_hitlerum"
}
```

Figure 1: Our proposed LLM output for a reduced input tweet from the HQP dataset (Maarouf et al., 2023) where it was initially weak-labeled as "slogans".

1 Introduction

Fake news and disinformation have become a significant challenge, particularly in geopolitical conflicts like the Russia-Ukraine war (Perez, 2022). Disinformation campaigns strategically manipulate public opinion and shape narratives (Wardle and Derakhshan, 2017; Zhdanova and Orlova, 2017), with pro-Russian biases linked to reduced ability to identify propaganda (Erlich and Garner, 2023). Propaganda, defined as *"the deliberate and systematic attempt to shape perceptions, manipulate cognitions, and direct behavior to achieve a response that furthers the desired intent of the propagandist"* (Lock and Ludolph, 2020; Jowett and O'donnell, 2018), lies at the core of these campaigns. Detecting such manipulative content is critical for preserving public trust and safeguarding democratic processes (Bayer et al., 2021). While propaganda in long-form text is well studied (Martino et al., 2020), short-form propaganda remains more challenging due to limited annotated data, sparse context, and the use of informal language,

* Equal contribution

¹https://github.com/XplainNLP/NLP4PI_2025_submission

abbreviations, and hashtags (Vijayaraghavan and Vosoughi, 2022). Although automated methods for disinformation and propaganda detection have advanced (Plikynas et al., 2025), the task remains difficult. Subtle linguistic cues, context-dependent interpretations, and low inter-annotator agreement highlight the complexity of human annotations (Hasanain et al., 2023; Srba et al., 2024), particularly in fine-grained classification (Hasanain et al., 2024; Martino et al., 2020), as propaganda often exploits cognitive biases and undermines critical thinking, making individuals more susceptible to conspiratorial narratives (Tanvir and Malik, 2024; Sahitaj et al., 2024). Propaganda detection aligns with the United Nations Sustainable Development Goal (SDG) 16², which promotes peaceful, inclusive societies and effective institutions. Misinformation and propaganda undermine these aspirations by fueling social divisions, eroding trust in institutions, and obstructing transparent communication (Mwangi, 2023), especially when amplified by automated bots (Zhdanova and Orlova, 2017). In this work, we propose a methodology that advances propaganda detection through the following five key contributions: First, we develop a fine-grained propaganda taxonomy that categorizes 14 distinct techniques by Martino et al. (2020) into three broader groups based on their intent: those that trigger emotional responses, those that simplify or distort complex issues, and those that undermine trust through authority and group dynamics. Second, we conduct an initial human annotation study on a statistically significant subset of propagandistic tweets from the HQP dataset (Maarouf et al., 2023). This study highlights the challenges of manual fine-grained labeling, revealing that the process is highly subjective, time-consuming, and prone to low inter-annotator agreement. Third, to overcome these limitations, we propose a novel LLM-assisted annotation methodology. In our pipeline, LLMs first extract relevant propaganda spans from the text, explain why these spans are considered propagandistic, and then assign fine-grained labels at the span level before determining a global label for the entire post. Fourth, we perform a secondary human verification study on a stratified sample of LLM-annotated posts. In this stage, human annotators are presented with the extracted spans and their local labels, and tasked with annotating the global propaganda label. We observe that annota-

tion agreement increases, and time investment is reduced by introducing LLMs as pre-annotation tool. Finally, we fine-tune small language models on the LLM-generated annotations to perform structured span-based labeling and explanation, enabling scalable training through knowledge distillation without relying on human-labeled data.

2 Related Work

Early research on automatic propaganda detection approached the problem at the document level, aiming to classify entire news articles (Rashkin et al., 2017). For instance, some systems labeled texts into four broad categories (trusted, satire, hoax, or propaganda) (Rashkin et al., 2017), while others framed it as a binary task (propaganda, non-propaganda) (Barrón-Cedeno et al., 2019), which limited granularity and explainability (Martino et al., 2019). An advance came with the work of Martino et al. (2019), who introduced span-level analysis with the PTC corpus, which comprises news articles annotated at the sentence level and fragment level with 18 distinct propaganda techniques. This scheme was adopted by the SemEval-2020 Shared Task (Martino et al., 2020) which consolidated the 18 techniques into a set of 14 widely used labels (Martino et al., 2020; Sprenkamp et al., 2023; Abdullah et al., 2022), that we also follow in our work. Early models used BERT-based architectures to perform span identification and technique classification (Da San Martino et al., 2019). Building on this, recent work explores how LLMs can further enhance propaganda detection, in terms of reducing annotation time and cost while improving label agreement and quality across classification tasks (Alizadeh et al., 2025; Gilardi et al., 2023; Ding et al., 2022). However, the use of LLMs may also exhibit stronger systematic bias than human annotators, especially in politically sensitive contexts (Vera and Driggers, 2024), and may suffer from generation-related issues such as hallucinations (Lee, 2023). Within propaganda detection, Jose and Greenstadt (2025) evaluated GPT-3.5, GPT-4, and Claude on identifying six propaganda techniques in news articles. Hasanain et al. (2023) employed GPT-4 as an LLM-as-Annotator approach to annotate Arabic text spans with 23 propaganda techniques using multilabel and sequence tagging tasks, and trained BERT-based models on the generated annotations. Similarly, Sprenkamp et al. (2023) examined the performance of multiple

²<https://sdgs.un.org/goals/goal16>

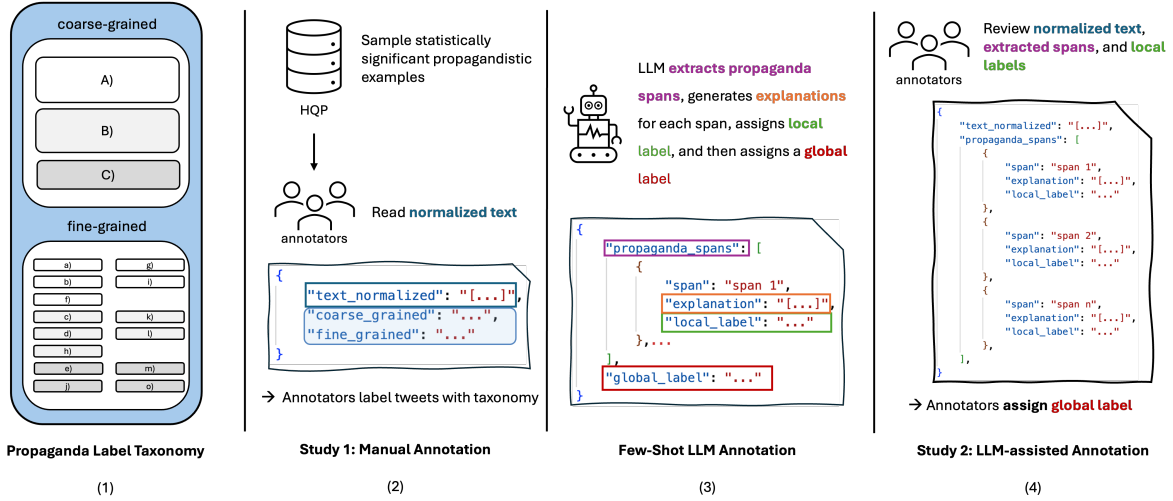


Figure 2: Methodological Overview

GPT-3 and GPT-4 variants for multi-label classification of 14 propaganda techniques at article-level using the SemEval-2020 Task 11 dataset (Martino et al., 2020), employing a range of prompt engineering and fine-tuning strategies. Their results show that GPT-4 can approach state-of-the-art performance. Our work builds on these efforts by grouping the 14 fine-grained techniques (Martino et al., 2020) into a novel coarse-grained taxonomy of three broader categories to support human annotator clarity and enable hierarchical modeling. By using a fully open-source LLM (LLaMA3-70B), we extract propaganda spans from tweets and assign fine-grained local labels based on the 14 techniques from Martino et al. (2020). In addition, it assigns a global propaganda label that captures the tweet’s overall framing. While the LLM also generates explanations for why each span was classified as propagandistic, these are not shown to human annotators but are used as an intermediate reasoning step to guide models towards their prediction. Moreover, we distill four small student models on the generated outputs of the larger model as teacher to enable propaganda span in resource-constrained environments through an open-source modeling pipeline.

3 Methodology and Results

In this Section, we outline our novel methodology that combines human expertise with computational techniques, as displayed in Figure 2, and their results. We first define a labeling framework for both coarse-grained and fine-grained categories in Section 3.1. Next, we describe our human annotation

study (Study 1, see Section 3.2) on the HQP dataset (Maarouf et al., 2023). We then detail our LLM few-shot inference in Section 3.3 and annotation approach, to automatically extract propaganda spans, generate explanations, and assign fine-grained labels, followed by a second human verification study (Study 2, see Section 3.4). Finally, we fine-tune SMLs via knowledge distillation in Section 3.5.

3.1 Propaganda Label Taxonomy

Annotating text for propaganda techniques is a highly complex task, as it is influenced by subjectivity, cognitive biases, personal experiences and the subtle variations in meaning that arise from different cultural and linguistic contexts (Sprenkamp et al., 2023). Prior work has highlighted that distinguishing between multiple fine-grained techniques can be particularly demanding, leading to low inter-annotator agreement and making it difficult to maintain consistency across annotations. (Hasanain et al., 2024)

To investigate this problem, we survey the literature and aggregate definitions from previous works, most notably the 14 propaganda techniques introduced by Martino et al. (2020), which refined an earlier set of 18 techniques proposed by Martino et al. (2019) and later applied by Sprenkamp et al. (2023) and Abdullah et al. (2022) to analyze and label propaganda techniques in text.

In our framework, the fine-grained propaganda techniques are organized in broader, coarse-grained categories according to their manipulative intent and rhetorical function. Detailed definitions of the techniques can be found in the Appendix A.1. This hierarchical framework aims to reduce cognitive

load for annotators and improve labeling consistency by first categorizing propaganda into conceptual groups before applying fine-grained classifications. It also enables us to evaluate the fine-grained predictions within the context of the coarse-grained labeling system in the subsequent analysis. The three coarse-grained categories are as follows:

- (A) **Emotional Appeals to Influence Opinions and Behaviors.** Techniques that exploit emotions to influence opinions or actions, often by-passing rational analysis. These methods use emotionally charged language, imagery, or ideas to evoke strong feelings. It includes the following techniques: loaded language, name calling, labeling, appeal to fear/prejudice, flag-waving, slogans.
- (B) **Simplification and Distortion Strategies.** Techniques that distort reality by presenting complex issues in oversimplified or misleading ways. These methods often aim to reduce critical thinking and encourage binary or superficial understanding. Here, the following techniques are included: repetition, exaggeration or minimization, causal oversimplification, black-and-white fallacy, thought-terminating clichés.
- (C) **Manipulating Trust, Authority, and Rational Discourse.** Techniques that undermine trust, exploit authority, discredit opponents, or manipulate group dynamics to shift opinions. These methods often redirect attention or leverage associations to influence perceptions of credibility or legitimacy. This includes the following techniques: doubt, appeal to authority, whataboutism, straw man, red herring, bandwagon, reductio ad hitlerum.

3.2 Study 1: Human Annotation

In this initial study, we aim to replicate previous findings from Hasanain et al. (2024) that emphasize the challenges of annotating fine-grained propaganda techniques, most notably, the low inter-annotator agreement (IAA) observed in such tasks. For studying the annotation of fine-grained labels, we utilize the HQP dataset (Maarouf et al., 2023), which comprises 29,596 tweets annotated for binary propaganda detection within the context of Russian propaganda. Out of these, 4,534 tweets were previously identified as propagandistic.

Assuming that the binary classification of propaganda versus non-propaganda is reliable, we confined our analysis to the subset of tweets labeled as propaganda. This focus allowed us to isolate the task of assigning detailed, fine-grained labels without the confounding effects of binary misclassification. Based on a 5% margin of error at a 95% confidence level and following established sample size estimation methods (Ahmed, 2024), a sample of $n = 355$ was selected from the 4,534 tweets labeled as propaganda. While this sample is statistically sufficient to estimate proportions, we consider this a pilot study to explore annotation feasibility and qualitative patterns rather than claiming full representativeness of the corpus.

3.2.1 Setup

Initially, the annotators were provided with the HQP annotation guidelines (Maarouf et al., 2023), which define propaganda as *deliberate expressions aimed at influencing opinions*, with a specific focus on Russian propaganda in the context of the Russo-Ukrainian conflict. This ensured a common understanding of the binary classification of tweets as propagandistic. Subsequently, they received a supplementary annotation guideline that included the previously introduced definitions and concrete examples of both coarse-grained and fine-grained propaganda categories. Annotators were instructed to first select the most appropriate coarse-grained category and then assign the single most significant fine-grained label for each tweet.

3.2.2 Results

The first human annotation study required three annotators to label each tweet using both the pre-defined coarse-grained categories and the more detailed fine-grained labels. The coarse-grained labels achieved a moderate level of consensus, as seen in Table 1.

Metric	Coarse	Fine
Raw Agreement 2/3	0.8845	0.4761
Raw Agreement 3/3	0.2789	0.0761
Krippendorff's Alpha	0.2065	0.1233

Table 1: Inter-Annotator Agreement Metrics for Coarse- and Fine-Grained Propaganda Annotations in Round 1.

Specifically, the raw agreement for coarse-grained annotations reached 88.45% with a 2/3 majority but dropped to 27.89% when full 3/3 consensus was required. The fine-grained labeling

presented greater challenges, with the raw agreement (2/3) being 47.61%, while the full agreement reached only 7.61%. The corresponding Krippendorff’s Alpha values of coarse and fine-grained labels further underscore the limitations in obtaining consistent fine-grained annotations. A more detailed analysis in Table 2 reveals that fine-grained agreement improves substantially when annotators already agree on the coarse-grained category.

Subset	2/3 Fine	3/3 Fine
2/3 Coarse	0.4372	0.0000
3/3 Coarse	0.7475	0.2727

Table 2: Fine-grained agreement rates conditioned on prior majority 2/3 or full 3/3 agreement on coarse labels.

In the guidelines for the annotation of the HQP dataset (Maarouf et al., 2023), annotators were asked to label the entire tweet as propagandistic, even if only some segments of the text contain propagandistic content. While we followed this notion for our own annotation of fine-grained labels, our analysis revealed that many tweets comprised multiple segments, each potentially associated with different propaganda labels. This complexity made applying a single definite label to the entire document challenging, as annotators not only had to differentiate among 14 possible labels but also rank the labels based on their impact, so that they could choose the most prominent one. This additional layer of subjectivity and specificity, also contributing to an average annotation time of 151.70 seconds per instance, underscores the need to explore alternative annotation strategies, such as LLM-assisted pre-annotation, as discussed in the following sections.

3.3 Few-Shot LLM Annotation

Based on the findings of Study 1, we extend the annotation approach by implementing an LLM to extract segments of potential propagandistic content and assign labels at two levels. In this approach, the LLM is tasked with three subtasks: (i) extracting spans from the presented tweet that likely contain propagandistic language, (ii) generating concise explanations for why each span was classified as propagandistic, and (iii) assigning a fine-grained local label to each extracted span as well as a global label for the entire tweet.

We employ few-shot inference with llama3.3-70B-Instruct model (AI@Meta, 2024). Specifically, we

create a synthetic few-shot example for each of the fine-grained propaganda labels and incorporate the corresponding label definitions into the system prompt. Each example is manually constructed to reflect a typical use of the respective technique. Three of the authors review each example for clarity and fit. We utilize structured generation to ensure that outputs can be easily parsed and evaluated (Willard and Louf, 2023). No additional background knowledge about the content of the situation is provided, so that the LLM relies solely on the few-shot examples and the label definitions to perform the task. The prompt is presented in the Appendix A.4 in Figure 7 and 8.

3.3.1 Results

The LLM was applied to all tweets labeled as propagandistic in the HQP dataset (Maarouf et al., 2023). In 94 cases, the model did not detect any propagandistic span. Upon manual analysis, we identified that 30 of these cases did exhibit rather clear propagandistic technique or framing. However, without specific contextual knowledge, these cases could often be mistaken for opinion pieces or news. The remaining majority were news reports, discussions, or opinion pieces that did not include explicit propaganda. For the following analysis, we filtered out these cases.

The distribution of predicted global labels is summarized in the Appendix A.2 in Table 8. The most common labels were loaded_language, doubt, reductio_ad_hitlerum, and name_calling. Prior work has noted that reductio_ad_hitlerum is a frequent technique in Russian propaganda (Gherasim, 2022). In our setting, this label appears alongside similar categories such as loaded_language and name_calling, suggesting empirical overlaps in how these techniques are used. Next, we examined the number of detected propaganda spans per tweet (the distribution is illustrated in Table 3). Our empirical results suggest that a majority of the propagandistic tweets contain multiple propagandistic segments. Relying solely on assigning a global label as has been focused by previous work, may therefore lead to a loss of important details, indicating that future work should maintain the extraction of segments and their local labels as primary target.

spans	1	2	3	4	5+
count	289	1,119	1,663	1,002	367

Table 3: Distribution of detected Propaganda Spans.

Focusing on tweets with at least three extracted propaganda spans, which is 3,032 cases, we observed that in 76.65% of these instances, the local label assigned to the first extracted span matched the global label for the entire tweet. This suggests a strong tendency for the most impactful propagandistic content to appear at the beginning of tweets. Furthermore, about 30% of cases with at least three extractions, exhibited a majority of local labels. In 83.55% of these cases, this majority local label also aligned with the global label. Thus, we observe that the dominant propaganda technique can be inferred when a majority of extracted local labels is available.

3.3.2 Ablation

To assess the robustness of our approach, we conducted several ablation studies. In the first analysis, we compared tweet annotations generated from normalized text (i.e., text with usernames, links, and similar elements removed) against those from non-normalized tweets. To statistically evaluate the differences between these paired categorical observations, we employed the Stuart-Maxwell (marginal homogeneity) test. Under the null hypothesis H_0 , that the proportion for each predicted global label in the normalized variation is equal to that of the original tweet text. The Stuart-Maxwell test yields a test statistic of 15.32 with 16 degrees of freedom and a p -value of 0.5014. Consequently, we conclude that there is no significant difference between the annotated global labels obtained from normalized versus non-normalized text.

Next, we evaluated the stability of the LLM’s outputs by repeating the experiment $k = 5$ times. Initially, under standard conditions with static few-shot examples, consistent task descriptions, and guided decoding, our approach yielded stable results for the extracted spans, assigned local labels, and the global label in 5/5 cases. To further challenge the model’s robustness, we introduce maximum randomness by shuffling the order of the few-shot examples and the label definitions in the prompt for each data point. We noted the agreement across five runs, randomized for each data point in each run (Table 4). These results indicate that even under maximum prompt randomness, our approach remains quite robust. Nonetheless, variations in the ordering of few-shot examples and label definitions have a marginal effect, particularly on local label predictions, whereas the extracted spans and global label predictions remain more stable.

This observation reinforces our initial finding that certain extracted spans may correspond to multiple appropriate labels while still being associated with a consistent global label.

Agreement	$\geq 3/5$	$\geq 4/5$	5/5
Local Label	100.00%	95.46%	81.48%
Extract. Spans	100.00%	97.74%	89.86%
Global Label	100.00%	98.58%	94.17%

Table 4: Agreement across 5 runs with randomization.

3.4 Study 2: Human Annotation

In this second human annotation study, we aim to assess whether integrating LLM-generated annotations with human verification improves annotation consistency and efficiency. Unlike the first study, where annotators assigned coarse- and fine-grained labels without assistance, this study provides them with LLM generated pre-annotations as optional suggestions. Annotators are presented with the original normalized tweet, the extracted spans, and corresponding labels, but they do not modify or verify individual spans. Instead, they select the most appropriate coarse-grained category and fine-grained technique for the entire tweet from a pre-defined set of options. The predicted global label of the LLM remains hidden while annotating, ensuring that human decisions are less biased and independent of the model’s final classification. To minimize potential bias from task familiarity, we exclude the most experienced annotator and swap them with an annotator who has not participated in the first study. This approach is intended to introduce a regularization effect and ensure a more balanced evaluation.

3.4.1 Setup

The annotation process in this study followed the same structured approach as described in the setup in Section 3.2.1. However, instead of selecting tweets randomly, we employed a stratified sampling approach based on the global labels predicted by the LLM. Since the distribution of propaganda techniques in real-world data is often imbalanced, random sampling could result in over-representation of some categories and under-representation of others. To ensure that each global label was sufficiently covered, we stratified the sample according to the LLMs predicted global propaganda labels. Most global labels predicted by the LLM appeared frequently in the dataset, allowing for an even allocation across categories.

However, techniques such as bandwagon and repetition were considerably less prevalent in the full dataset of 4,534 propagandistic tweets, occurring only 8 times and 6 times, respectively. Based on that, all occurrences of these global labels were included in the sample to ensure that they were adequately represented in the analysis.

3.4.2 Results

In the second human annotation study, annotators were provided with LLM-generated pre-annotations that include extracted propagandistic spans along with corresponding local fine-grained labels. However, the predicted global label by the LLM was not shown to them, and annotators remained fully responsible for independently selecting the global coarse- and fine-grained label for each tweet. Compared to Study 1, this approach led to notable improvements in IAA as well as annotation efficiency.

Metric	Coarse	Fine
Raw Agreement 2/3	0.9746	0.9014
Raw Agreement 3/3	0.6225	0.4789
Krippendorff's Alpha	0.6059	0.5941

Table 5: Inter-Annotator Agreement Metrics for Coarse- and Fine-Grained Propaganda Annotations in Round 2.

As shown in Table 5, the raw agreement for coarse-grained labels increased from 88.45% (2/3 majority) and 27.89% (full consensus) in Study 1 to 97.46% (2/3 majority) and 62.25% (full consensus) in Study 2. For fine-grained labels, raw agreement improved from 47.61% 2/3 and 7.61% 3/3 in Study 1 to 90.14% 2/3 and 47.89% in Study 2, respectively. Correspondingly, the Krippendorff's Alpha increased from 0.2065 (coarse) and 0.1233 (fine) in Study 1, to 0.6059 (coarse) and 0.5941 (fine) in Study 2. A detailed examination of fine-grained agreement rates conditioned on the level of consensus in the coarse labels in Table 6 further confirms these improvements. In Study 2, these rates improved to 80% for tweets with a (2/3) coarse consensus, and for tweets with full coarse consensus, the (2/3) fine-grained agreement increased to 99.55%, with full (3/3) fine-grained agreement at 76.02%. An illustrative example of the effectiveness of LLM-assisted annotation is shown in Figure 3. In this instance, the LLM successfully identified key propagandistic spans, assigned appropriate fine-grained labels, and provided coherent explanations that aligned well with human interpretations. In this case, the hashtag

```
{
  "text_normalized": "#IStandWithPutin Bcoz Russia is
our true friend. HTTPURL",
  "propaganda_spans": [
    {
      "span": "#IStandWithPutin",
      "explanation": "Uses a hashtag to express
support for Putin and create a sense of
solidarity.",
      "local_label": "slogans"
    },
    {
      "span": "Russia is our true friend",
      "explanation": "Frames Russia as a loyal and
trustworthy ally to garner emotional support.",
      "local_label": "flag-waving"
    }
  ],
  "global_label": "slogans"
}
```

Figure 3: Example of LLM-assisted annotation, showing accurate span extraction, fine-grained label assignment, and coherent explanations. This case achieved full 3/3 IAA.

"#IStandWithPutin" was labeled as slogans, reinforcing ideological solidarity, while "Russia is our true friend" was classified as flag-waving, portraying Russia as a trustworthy ally. The explanations clearly justify the propagandistic nature of each span, and the global label ("slogans") is particularly suitable, as slogans, especially when used as hashtags, are concise and easily shareable, amplifying their spread on social media and reinforcing group identity more effectively than descriptive statements. This annotation achieved full 3/3 IAA, confirming its reliability.

Subset	2/3 Fine	3/3 Fine
2/3 Coarse	0.8000	0.0160
3/3 Coarse	0.9955	0.7602

Table 6: Fine-grained agreement rates conditioned on prior majority 2/3 or full 3/3 agreement on coarse labels in Round 2.

Additionally, Cohen's Kappa was calculated to measure agreement between human majority-vote labels and LLM-generated global labels. If no 2/3 majority was reached, a random LLM prediction was used as the human label. The resulting Cohen's Kappa score of 0.8438 indicates strong agreement between human annotations and LLM-generated global labels. Also, the average annotation time per tweet is reduced from 151.70 seconds in Study 1 to 41.14 seconds in Study 2. In summary, the integration of LLM-generated pre-annotations with human verification in Study 2 resulted in higher

IAA and reduces annotation time relative to the fully manual approach in Study 1, indicating an overall improvement in reliability, efficiency and scalability.

3.5 Knowledge Distillation

Based on our findings, we next aim to scale structured propaganda annotation and enable efficient inference in resource-constrained environments by fine-tuning a collection of SLMs on LLM-generated supervision. In this knowledge-distillation-inspired setup, the 70B model as described in Section 3.3 serves as the *teacher*, providing structured propaganda annotations for every data point. We train four *student* models, two LLaMA3-based variants (3B and 8B parameters) denoted as *L*, and two Qwen2.5 variants (3B and 7B parameters) denoted as *Q*. To minimize memory usage and accelerate training, we employ parameter-efficient fine-tuning (PEFT), combined with 4-bit quantization. We employ a standard sequence-to-sequence cross-entropy loss, without additional regularization terms or explicit teacher-student logit matching, to generate the structured responses. We utilize a stratified 80/20 split and learn on the train split for three epochs.

3.5.1 Results

We report six evaluation metrics on the unseen test set as reported in Table 7. Here, **G** denotes the macro- and micro-averaged global F1 scores over the test set. **Span_e** describes the F1 for exact span detection, while **Span_f** specifies the fuzzy-span F1 with a strict 0.8 similarity threshold to account for minor variations following the notion of partial matches as introduced by (Hasanain et al., 2023). Similarly, **Local_e** requires both exact span text and correct local label classification, while **Local_f** combines fuzzy span matching with correct local label assignment.

Model	G_{macro}	G_{micro}	Span_e	Span_f	Local_e	Local_f
<i>L_{3b}</i>	0.49	0.36	0.40	0.60	0.22	0.32
<i>L_{8b}</i>	0.58	0.47	0.47	0.67	0.29	0.40
<i>Q_{3b}</i>	0.48	0.34	0.40	0.61	0.21	0.31
<i>Q_{7b}</i>	0.51	0.34	0.45	0.66	0.25	0.36

Table 7: Student Model Evaluation Results.

All four student models achieve reasonable performance on each metric. Larger models show modest gains, and *L* and *Q* variants of the same size perform similarly. Global-label prediction across 14 propaganda categories (Martino et al.,

2020) yields acceptable F1 scores, suggesting that choosing a global label is relatively straightforward. Span detection also works well under both exact-match and fuzzy-match criteria. By contrast, assigning local labels remains difficult. Models reliably find propaganda spans but are less certain which specific technique to annotate. We hypothesize that this stems from two key factors: (1) the limited volume of training data available for fine-grained local label predictions, and (2) the inherent ambiguity due to overlap in the definitions of certain propaganda techniques, while the general notion of a propaganda span seems to be more solid.

4 Discussion and Conclusion

In this paper, we introduced an LLM-assisted annotation framework that combines automated extraction of propaganda spans with human verification. Our experiments demonstrate that integrating LLM-assisted pre-annotation with human verification significantly improves the consistency and efficiency of propaganda detection. In Study 1, manual fine-grained labeling suffered from low inter-annotator agreement and long annotation times. Study 2, which incorporated LLM-generated pre-annotations based on extracted propaganda spans, yielded higher agreement metrics and reduced annotation time, although part of the efficiency gain may stem from annotators’ familiarity with the task. Notably, our results suggest that a single global label is sometimes insufficient to capture the complexity of propagandistic content, as our analysis shows most tweets include more than one extracted propaganda span. This granular perspective may offer better insights than traditional sequence-level classification, and it is more scalable across different text lengths. These findings, in line with emerging trends such as those highlighted in SemEval-2023 Task 3 (Piskorski et al., 2023), indicate that future work should consider reformulating the problem to emphasize alternative propaganda detection strategies. Exploring multi-label and hierarchical annotation strategies may better accommodate the overlapping nature of propaganda techniques. Finally, integrating richer contextual information and real-time fact-checking modules could further refine detection performance (Sahitaj et al., 2025). We also advocate for iterative human-in-the-loop systems that continuously update few-shot examples and label definitions to minimize bias and enhance model robustness.

Limitations

While promising, our approach has several limitations. First, our study is confined to English tweets related to Russian propaganda which may limit its applicability to other languages or domains. Second, the reliance on a single global label despite the local span-based analysis might oversimplify instances where multiple propaganda techniques coexist. Third, some improvements in annotation efficiency could be attributed to annotator learning effects rather than solely to the LLM-assisted pre-annotation. Fourth, the quality of LLM-generated pre-annotations depends on the few-shot examples and definitions provided which could introduce bias or inconsistencies. Following work should involve a larger and more diverse pool of annotators to further validate and refine the framework. In addition, self-collected data from various propaganda settings encompassing multiple languages and platforms would offer a broader evaluation and help mitigate potential biases inherent in the current dataset. Another limitation concerns our distillation setup. Biases present in the 70B teacher model due to its pretraining may be propagated to the student models. Since the student models are trained solely on model-generated supervision any ideological or geopolitical bias in the teacher can persist without correction. While the use of open-source models improves transparency and auditability it does not inherently prevent bias propagation. Future work should systematically investigate inherited bias in open-source propaganda detection pipelines.

Ethical and societal implications

The integration of LLM-assisted annotation in propaganda detection raises ethical concerns regarding bias, automation dependency, misuse, and public trust. While improving annotation efficiency, LLM-generated labels may introduce systematic biases, reflecting dominant narratives in their training data. This can influence human annotators' decisions, leading to reinforced biases instead of neutral classifications. Another risk is automation bias, where annotators overly rely on LLM suggestions and reduce their critical thinking ability. Furthermore, such models could be exploited for counter propaganda, with governments or other actors potentially using them to suppress dissenting voices and shape public discourse to their advantage. Faulty or overly simplistic propaganda detection may inad-

vertently weaken trust in media and public institutions, undermining the democratic ideals promoted by SDG 16. Therefore, it is imperative that the development and deployment of these systems remain transparent, incorporate rigorous bias audits, and maintain robust human oversight to ensure that they support democratic discourse rather than restrict it.

Acknowledgments

This research is funded by the Federal Ministry of Research, Technology and Space (BMFTR, reference: 03RU2U151C) in the scope of the research project news-polygraph.

References

- Malak Abdullah, Ola Altit, and Rasha Obiedat. 2022. Detecting propaganda techniques in english news articles using pre-trained transformers. In *2022 13th International Conference on Information and Communication Systems (ICICS)*, pages 301–308. IEEE.
- Sirwan Khalid Ahmed. 2024. How to choose a sampling technique and determine sample size for research: a simplified guide for researchers. *Oral Oncology Reports*, 12:100662.
- AI@Meta. 2024. [Llama 3 model card](#).
- Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan D Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2025. Open-source llms for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science*, 8(1):1–25.
- Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Propopy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.
- Judit Bayer, Bernd Holznagel, Katarzyna Lubianiec, Adela Pintea, Josephine B Schmitt, Judit Szakács, and Erik Uszkiewicz. 2021. Disinformation and propaganda: impact on the functioning of the rule of law and democratic processes in the eu and its member states. *European Union*.
- Giovanni Da San Martino, Alberto Barron-Cedeno, and Preslav Nakov. 2019. Findings of the nlp4if-2019 shared task on fine-grained propaganda detection. In *Proceedings of the second workshop on natural language processing for internet freedom: censorship, disinformation, and propaganda*, pages 162–170.
- Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

- Aaron Erlich and Calvin Garner. 2023. Is pro-kremlin disinformation effective? evidence from ukraine. *The International Journal of Press/Politics*, 28(1):5–28.
- Gabriel C Gherasim. 2022. Reductio ad hitlerum: Reflections on the russian propaganda of de-nazification in ukraine. *Romanian Journal of Political Sciences*, 22(1):75–86.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024. [Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2724–2744, Torino, Italia. ELRA and ICCL.
- Maram Hasanain, Fatema Ahmed, and Firoj Alam. 2023. Large language models for propaganda span annotation. *arXiv preprint arXiv:2311.09812*.
- Julia Jose and Rachel Greenstadt. 2025. Are large language models good at detecting propaganda? *arXiv preprint arXiv:2505.13706*.
- Garth S Jowett and Victoria O’donnell. 2018. *Propaganda & persuasion*. Sage publications.
- Minhyeok Lee. 2023. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*, 11(10):2320.
- Irina Lock and Ramona Ludolph. 2020. Organizational propaganda on the internet: A systematic review. *Public Relations Inquiry*, 9(1):103–127.
- Abdurahman Maarouf, Dominik Bär, Dominique Geissler, and Stefan Feuerriegel. 2023. Hqp: a human-annotated dataset for detecting online propaganda. *arXiv preprint arXiv:2304.14931*.
- G Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. Semeval-2020 task 11: Detection of propaganda techniques in news articles. *arXiv preprint arXiv:2009.02696*.
- Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. *arXiv preprint arXiv:1910.02517*.
- Eric Mwangi. 2023. Technology and fake news: shaping social, political, and economic perspectives. *Political, and Economic Perspectives (May 29, 2023)*.
- E Perez. 2022. Strategic disinformation: Russia, ukraine and crisis communication in digital era.
- Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. [SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.
- Darius Plikynas, Ieva Rizgelienė, and Gražina Korvel. 2025. Systematic review of fake news, propaganda, and disinformation: Examining authors, content, and social impact through machine learning. *IEEE Access*.
- Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.
- Ariana Sahitaj, Premtim Sahitaj, Salar Mohtaj, Sebastian Möller, and Vera Schmitt. 2024. Towards a computational framework for distinguishing critical and conspiratorial texts by elaborating on the context and argumentation with llms. *Working Notes of CLEF*.
- Premtim Sahitaj, Iffat Maab, Junichi Yamagishi, Jawan Kolanowski, Sebastian Möller, and Vera Schmitt. 2025. [Towards Automated Fact-Checking of Real-World Claims: Exploring Task Formulation and Assessment with LLMs](#). *Preprint*, arXiv:2502.08909.
- Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*.
- Ivan Srba, Olesya Razuvayevskaya, João A Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, et al. 2024. A survey on automatic credibility assessment of textual credibility signals in the era of large language models. *arXiv preprint arXiv:2410.21360*.
- Muhammad Tanvir and Azeem Malik. 2024. The information battlefield: How cyber propaganda affects thoughts and shape the public opinion. *Wah Academia Journal of Social Sciences*, 3(2):258–279.
- Sebastián Vallejo Vera and Hunter Driggers. 2024. Bias in llms as annotators: The effect of party cues on labelling decision by large language models. *arXiv preprint arXiv:2408.15895*.
- Prashanth Vijayaraghavan and Soroush Vosoughi. 2022. Tweetspin: Fine-grained propaganda detection in social media using multi-view representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448.
- Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework*

for research and policymaking, volume 27. Council of Europe Strasbourg.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*.

Mariia Zhdanova and Dariya Orlova. 2017. Computational propaganda in ukraine: Caught between external threats and internal challenges.

A Appendix

A.1 Fine-grained labels

The definitions of the propaganda techniques presented here are based on the 14 categories introduced by Martino et al. (2020), which refined an earlier set of 18 techniques proposed in Martino et al. (2019). These 14 categories have also been utilized in later works, such as Sprenkamp et al. (2023) and Abdullah et al. (2022), to analyze and label propaganda techniques in text.

- a) **Loaded language** involves the use of words or phrases with either strong positive or negative emotional connotations, to shape audience perceptions and influence their opinions.
- b) **Name calling, labeling** involves assigning a specific label to a target, intended to evoke either positive or negative emotions in the audience, such as fear, hatred, admiration, or praise.
- c) **Repetition** is the continuous repetition of a message or idea to increase its acceptance by the audience over time.
- d) **Exaggeration or minimization** involves portraying something in an overstated manner to amplify its significance or downplaying its importance to make it appear less impactful than it truly is.
- e) **Doubt** involves raising uncertainty or questioning the credibility of an individual, group, or entity to undermine trust.
- f) **Appeal to fear/prejudice** aims to built support for an idea by evoking anxiety, fear, or panic in the audience, often directed at an alternative or based on existing biases.
- g) **Flag-waving** involves appealing to strong feelings of national or group identity, such as those tied to race, gender, or political affiliation, to justify or promote an action, idea, or individual as representative of the entire group.
- h) **Causal oversimplification** involves attributing an issue to a single cause while disregarding its complexity or the presence of multiple contributing factors. This may also include assigning blame to an individual or group without adequately exploring the complexity of the issue.
- i) **Slogans** are concise and striking phrases that often incorporate labeling or stereotyping, serving as emotional or cognitive appeals to influence beliefs or perceptions.
- j) **Appeal to authority** involves asserting that a claim is true solely based on the support of an authority or expert, without providing additional evidence. This can also include cases where the referenced individual lacks genuine expertise but is still presented as authoritative.
- k) **Black-and-white fallacy** involves presenting two opposing options as the only possible choices, disregarding the existence of other alternatives. In its extreme form, referred to as dictatorship, the audience is explicitly directed toward a specific action, effectively eliminating all other options.
- l) **Thought-terminating cliches** are short, generic phrases designed to suppress critical thinking and meaningful discussion, often by providing oversimplified answers to complex issues or diverting attention from deeper exploration of a topic.
- m) **Whataboutism, straw man, red herring** combines three distinct techniques, which are frequently grouped together due to their relatively rare individual usage. *Whataboutism* undermines an opponents argument by accusing them of hypocrisy without addressing their claims directly. *Straw man* misrepresents or distorts an opponents position by substituting it with a weaker or exaggerated version that is easier to refute. *Red herring* diverts attention from the main argument by introducing irrelevant information or topics.
- o) **Bandwagon, reductio ad hitlerum** combines two techniques often discussed together due to their similar persuasive nature. *Bandwagon*

attempts to convince the audience to adopt an idea or action by emphasizing that "everyone else is doing it". *Reductio ad hitlerum* seeks to discredit an idea or action by associating it with groups or individuals disliked or despised by the audience.

A.2 Global Labels Distribution

Table 8 provides an overview of the distribution of global propaganda labels predicted by the model across the dataset. As shown, the most frequently occurring techniques include loaded_language, doubt, reductio_ad_hitlerum, and name_calling.

Table 8: Distribution of predicted Global Labels

Global Label	Count
loaded_language	1384
doubt	647
reductio_ad_hitlerum	641
name_calling	519
whataboutism	333
appeal_to_fear_prejudice	250
causal_oversimplification	160
exaggeration	150
flag-waving	122
appeal_to_authority	106
straw_man	54
red_herring	54
thought-terminating_cliches	35
slogans	29
black-and-white_fallacy	25
repetition	17
bandwagon	8

A.3 Examples

In the HQP dataset (Maarouf et al., 2023), weak labeling was used to classify certain propaganda techniques. The HQP dataset (Maarouf et al., 2023) initially weak-labeled this tweet in Figure 4) as *slogans*. However, a more detailed analysis of the text spans reveals the presence of multiple propaganda techniques, including *loaded language*, *exaggeration*, and *reductio ad Hitlerum*.

In the future, by integrating fact-checking into propaganda detection, we can complement existing labeling approaches and assess whether the claims being made have a factual basis. This is important because propaganda often spreads through misinformation, and weak labels alone do not verify

truthfulness. Fact-checking strengthens the detection process by distinguishing between persuasive rhetoric and outright disinformation, making it a necessary component for a more precise and reliable analysis of propaganda content. (Sahitaj et al., 2025)

Another instance of weak labeling challenges in propaganda detection is presented in Figure 5. This example was initially annotated as propaganda in a binary setting by human annotators in the HQP dataset (Maarouf et al., 2023). In a later refinement, it was weak-labeled as slogans, reinforcing the classification as propagandistic content. However, during our qualitative analysis, we identified this as a case where the original annotation might not be justified.

```
{
  "input": "Ukrainian soldiers, the Nazi Azov battalion are shooting
their own Ukrainian citizens in the streets with snipers from
apartment windows HTTPURL",
  "propaganda_spans": [
    {
      "span": "Ukrainian soldiers, the Nazi Azov battalion are
shooting their own Ukrainian citizens",
      "explanation": "Makes a specific, unverified claim of
violence to demonize Ukrainian forces.",
      "local_label": "loaded_language"
    },
    {
      "span": "the Nazi Azov battalion",
      "explanation": "Associates the Azov battalion with Nazism
to evoke strong negative emotions.",
      "local_label": "reductio_ad_hitlerum"
    },
    {
      "span": "shooting their own Ukrainian citizens in the
streets with snipers from apartment windows",
      "explanation": "Uses vivid, disturbing imagery to create
outrage without providing evidence.",
      "local_label": "exaggeration"
    }
  ],
  "global_label": "loaded_language"
}
```

Figure 4: The example highlights how fact-checking is essential, as the detected spans include unverifiable claims of violence ('loaded language'), historical misrepresentation ('reductio ad hitlerum'), and exaggerated imagery ('exaggeration'). Without verification, such statements can contribute to misinformation and manipulation of public perception.

For our analysis, we specifically examined examples that were initially labeled as propaganda before receiving weak labels. This example was among them, but upon closer inspection, we do not find clear propagandistic intent. Instead, the text appears to be an analytical reflection or an ironic commentary on an existing narrative. This case highlights the difficulty of distinguishing between genuine propaganda and discourse that critically engages with a narrative. Automated or weak-labeling approaches may misclassify content

that shares linguistic patterns with propaganda but serves a different communicative function. These findings emphasize the need for more nuanced annotation approaches that incorporate contextual understanding, ensuring that content is not misclassified based solely on surface-level textual features. Another example illustrating the complex-

```
{
  "input": "logic of the events now: 1. Russian media has been telling its citizens for years that Ukraine is aggressive state 2. Intensified this narrative in past months 3. Says Ukraine makes military buildup, not Russia 4. Moscow taken by surprise by West's reaction, evacuations etc.",
  "propaganda_spans": [
    {
      "span": "Russian media has been telling its citizens for years that Ukraine is aggressive state",
      "explanation": "Creates a long-standing narrative to shape public opinion about Ukraine.",
      "local_label": "repetition"
    },
    {
      "span": "Intensified this narrative in past months",
      "explanation": "Increases the frequency or intensity of the message to reinforce the narrative.",
      "local_label": "repetition"
    },
    {
      "span": "Says Ukraine makes military buildup, not Russia",
      "explanation": "Reverses the reality of military buildup to shift blame.",
      "local_label": "causal_oversimplification"
    },
    {
      "span": "Moscow taken by surprise by West's reaction, evacuations etc.",
      "explanation": "Presents Russia as the victim, implying the West's reaction is unwarranted or disproportionate.",
      "local_label": "loaded_language"
    }
  ],
  "global_label": "repetition"
}
```

Figure 5: An example initially annotated as propaganda and weak-labeled as 'slogans' in the HQP dataset (Maarouf et al., 2023). During qualitative analysis, we found that this example does not necessarily exhibit clear propagandistic intent.

ity of propaganda detection is shown in Figure 6. This statement was initially annotated as propaganda in the HQP dataset (Maarouf et al., 2023) and subsequently relabeled using an LLM-based approach. The LLM did not perform binary classification but instead assigned fine-grained rhetorical labels, identifying thought-terminating clichés, red herring, and loaded language. However, during a qualitative review, we noticed that this example raises important questions about what should and should not be considered propaganda. This case is particularly interesting because, while the statement contains anti-Western criticism, which can be used in propaganda, it can also appear in normal discussions and political discourse. It does not necessarily display clear manipulative intent, even though it employs rhetorical techniques often associated with propaganda. The key challenge here is

```
{
  "input": "Russian propaganda isn't in any way sophisticated. It works because enormous sections of western (and Muslim) societies are in such deep cultural crisis that they have no respect for truth.",
  "propaganda_spans": [
    {
      "span": "Russian propaganda is n't in any way sophisticated",
      "explanation": "Downplays the complexity of Russian propaganda tactics.",
      "local_label": "thought-terminating_cliches"
    },
    {
      "span": "enormous sections of western (and Muslim) societies are in such deep cultural crisis",
      "explanation": "Suggests that the problem lies within the societies rather than addressing the propaganda itself.",
      "local_label": "red_herring"
    },
    {
      "span": "they have no respect for truth",
      "explanation": "Implies a broad lack of integrity in Western and Muslim societies.",
      "local_label": "loaded_language"
    }
  ],
  "global_label": "thought-terminating_cliches"
}
```

Figure 6: Initially annotated as propaganda and 'loaded language' in the HQP dataset (Maarouf et al., 2023) and later relabeled using an LLM. While it contains anti-Western criticism, such arguments can exist in normal discourse as well and are not a clear indicator of propaganda.

that rhetorical strategies alone do not automatically make a statement propagandistic. Context and intent matter. During our qualitative investigation of tweets, we found this to be a borderline case where one could argue both for and against labeling it as propaganda. On the one hand, its framing could serve as a tool for propaganda by reinforcing narratives about Western decline. On the other hand, such critiques exist independently of propaganda efforts. This example is valuable because it demonstrates that the LLM correctly assigned rhetorical strategies without overgeneralizing the statement as propaganda, highlighting the difficulty of drawing a clear boundary between manipulative content and critical discussion.

A.4 Prompts

The prompt establishes a structured framework for LLM-assisted annotation in propaganda detection, defining a systematic approach for identifying, explaining, and categorizing propagandistic content. As shown in Figures 7 and 8, the assistant is designed to extract specific spans indicative of propaganda, provide justifications based on predefined classification criteria, and assign both fine-grained local labels and an overarching global label. The framework (Figure 7) first guides the assistant to

detect key propaganda spans, classify them based on a predefined set of propaganda techniques, and explain why each span should be considered propaganda.

Prompt

SYSTEM:

You are an intelligent annotation assistant specializing in detecting propaganda. Your task is to analyze, explain, and pre-annotate the presented text based on a set of potential propaganda classifications. You **MUST** return the output in valid JSON following the defined schema.

****Setting**:** Detection of propaganda that is against the main opposition (i.e., Ukraine), against other oppositions (e.g., Western countries), or in favour of the Russian government.

1. ****Identify specific words or text spans that indicate propaganda.****
2. ****Explain for each extracted span why it should be considered propaganda.****
3. ****For each span, determine the dominant propaganda technique from the following list**:**
 - Loaded language: ...
 - Name calling: ...
 - Appeal to fear/prejudice: ...
 - Flag-waving: ...
 - Slogans: ...
 - Repetition: ...
 - Exaggeration/minimization: ...
 - Causal oversimplification: ...
 - Black-and-white fallacy: ...
 - Thought-Terminating Cliches: ...
 - Doubt: ...
 - Appeal to authority: ...
 - Whataboutism: ...
 - Straw man: ...
 - Red herring: ...
 - Bandwagon: ...
 - Reductio ad hitlerum: ...
4. ****Finally, assign the global label of the span that is most representative for the full sequence.****

Figure 7: Prompt (Part 1): Initial instructions for the propaganda detection task, including span extraction, explanation, and classification of local and global labels.

The second part (Figure 8) extends this process by enforcing a structured JSON output format, ensuring consistency across annotations and facilitating integration with human verification workflows. By structuring the annotation process in this way, our approach aims to improve labeling efficiency, reduce inter-annotator variability, and enhance the scalability of propaganda detection in large-scale datasets. The explicit categorization of rhetorical techniques provides a more detailed understanding of how propaganda manifests in text, while the standardized output format ensures that annotations remain interpretable and reproducible.

```
**Output Format**
Respond in **valid JSON** with the structure:
{
  "$defs": {
    "FineLabelVerdict": {
      "description": "Fine-grained categorization of
        propaganda techniques.",
      "enum": [
        ${LABELS}
      ]
    },
    "PropagandaSpan": {
      "description": "An identified propaganda span
        within the original text with an explanation.",
      "properties": {
        "span": {
          "description": "The exact propaganda span
            extracted from the original text.",
          "title": "Span",
          "type": "string"
        },
        "explanation": {
          "description": "The explanation why this
            span is considered propaganda.",
          "title": "Explanation",
          "type": "string"
        }
      },
      "local_label": {
        "$ref": "#/$defs/FineLabelVerdict",
        "description": "The appropriate label
          assigned towards the detected label."
      }
    },
    "required": [
      "span",
      "explanation",
      "local_label"
    ]
  },
  "global_label": {
    "$ref": "#/$defs/FineLabelVerdict",
    "description": "The label for the dominant
      propaganda technique in the statement."
  }
},
"description": "Schema for structured LLM output after
  propaganda detection and normalization."
}
USER:
${TWEET}

ASSISTANT:
```

Figure 8: Prompt (Part 2): JSON output format definition for our propaganda detection task.