Large Language Models for Propaganda Span Annotation

Anonymous ACL submission

Abstract

The use of propagandistic techniques in online content has increased in recent years, aiming to manipulate online audiences. Although essential for more informed content consumption; very limited focus has been given to the task of extracting textual spans where propaganda techniques are used. Our study focuses on that task by investigating whether large language models (LLMs), such as GPT-4, can effectively extract these spans. We further study the potential of employing the model to collect more cost-effective annotations. Our experiments use a large-scale in-house manually annotated dataset. The results suggest that providing more annotation context to the model as prompts improves its performance compared to human annotations. Moreover, our work is the first to show the potential of utilizing LLMs to develop annotated datasets for this complex task, prompting it with annotations from human annotators with limited expertise. All annotations will be shared with the community.[1]

1 Introduction

Malicious actors are actively exploiting online platforms to disseminate misleading content for political, social, and economic agendas (Perrin, 2015; Alam et al., 2022a; Sharma et al., 2022). The objective of using propaganda is to generate distorted and often misleading information, which can result in heightened polarization on specific issues and division among communities. Hence, it is important to automatically detect and debunk propagandistic content. The majority of relevant research has focused on either binary or multiclass and multilabel classification scenarios of the task (Barrón-Cedeno et al., 2019; Rashkin et al., 2017; Piskorski et al., 2023b). Very few studies have tackled the task of detecting propagandistic text spans (Da San Martino et al., 2019; Przybyła and Kaczyński, 2023). Span-level propaganda extraction in its nature is a complex task, as reported in different studies (Martino et al., 2020). This complexity is magnified by the large number of propaganda techniques that might be present (18 (Da San Martino et al., 2019) vs. 23 (Przybyła and Kaczyński, 2023) techniques for example). The subjective nature of the task also results in added challenges.

Several recent studies have benchmarked the capabilities of LLMs for downstream NLP tasks showing their remarkable capabilities (Bang et al., 2023; Ahuja et al., 2023; Abdelali et al., 2024a; Liang et al., 2022). However, the utility of LLMs in span-level propaganda detection remains underexplored. Therefore, we aim to leverage LLMs, selecting the most effective one to date, GPT-4 (OpenAI, 2023), for the task. Moreover, LLMs have shown to be effective aids in creating annotated datasets to train or evaluate other models in a variety of tasks (Alizadeh et al., 2023). Since there are many propaganda techniques to label and a need to create large and diverse datasets to train specialized models, LLMs might also benefit the process of developing new datasets for propaganda span detection. Recruiting humans to carry such large-scale annotations has been a very tedious and costly procedures. Thus, our study also aims to investigate whether we could use an LLM, such as GPT-4, to reduce the human annotation cost and effort by either reducing the number of annotators or hiring annotators with less expertise.

We study the following research questions: *(i)* Is GPT-4 capable of annotating spans effectively? *(ii)* Can GPT-4 serve both as a general and as an expert annotator of propaganda spans?[2] *(iii)* Which propaganda techniques can GPT-4 annotate best?

[1] to be made available

[2] For this task, the manual annotation process followed generally has two phases: *(i)* annotation done by three *general* annotators, *(ii)* annotations reviewed and disagreements resolved by two expert annotators. We use the term *general* to refer to less experienced but trained annotators.

The contributions of our study are as follows:

- We explore the use of GPT-4 as an annotator for detecting and labeling spans with propagandistic techniques, which is the *first attempt* at such a task. Results reveal the great potential of the model to replace more expert annotators, for some propaganda techniques (e.g., loaded language) with 36% reduction in annotation cost. We also provide an in-depth analysis of the model performance at different annotation stages.
- We are releasing annotations from human annotators and GPT-4 to benefit the community.

2 Related Work

Propaganda Detection. Relevant research has employed diverse methods to identify propagandistic text, ranging from analyzing content based on writing style and readability features in articles (Rashkin et al., 2017; Barrón-Cedeno et al., 2019) to using transformer based models for classification at the binary, multiclass and multilabel settings (Dimitrov et al., 2021b). Recent efforts stress the importance of fine-grained identification of specific propagandistic techniques. Da San Martino et al. (2019) identified 18 distinct techniques and created a dataset by manually annotating news articles based on them. Next, they designed a multi-granular deep neural network that extracts propagandistic spans from sentences with a limited F_1=22.58, showing how complex the task is. Piskorski et al. (2023b) extended the 18 techniques into 23 and introduced a dataset in multiple languages. With these efforts, fine-grained propaganda detection in general, and over Arabic content specifically, is still rarely investigated. Existing Arabic datasets are limited in size and number of targeted techniques (Alam et al., 2022b; Hasanain et al., 2023).

LLMs as Annotators. Constructing high-quality annotated datasets, essential for model training and evaluation, usually requires manual annotation by humans (Khurana et al., 2023). There has been efforts in utilizing LLMs for data annotation to overcome the challenges of human annotations, which include bias, time-overhead, and cost (Ding et al., 2023; Alizadeh et al., 2023; Thomas et al., 2023).

Sprenkamp et al. (2023) investigated the effectiveness of LLMs in annotating propaganda by utilizing five variations of GPT-3 and GPT-4. They tackled the task as a multi-label classification problem, using the SemEval-2020 Task 11 dataset. Their findings indicate that GPT-4 achieves results comparable to the current state of the art. Our work is closely related to theirs, however, they approached the problem as a multi-label text classification task of 14 techniques, at the article level. In contrast, we focus on fine-grained propaganda detection at the span level including both multilabel and sequence tagging tasks, covering 23 techniques, which is much more challenging.

3 Dataset

For this study, we utilized an in-house developed dataset. We briefly discuss the dataset development process. A complete detail of that process is beyond the scope of this paper.[3]

The dataset comprises annotated news paragraphs sourced from articles covering 300 Arabic news media. These news media are versatile, featuring a variety of writing styles and topics. It includes a total of **8,000 annotated paragraphs** selected from 2,800 news articles, approximately 10,000 sentences, and around 277,000 words. The dataset consists of 14 different topics, with 'news' and 'politics' accounting for over 50% of the paragraphs. The span level annotation agreement of the dataset is $\gamma = 0.546$. A brief of the annotation process is provided in Appendix A.

We split the dataset in a stratified manner (Sechidis et al., 2011), allocating 75%, 8.5%, and 16.5% for training, development, and testing, respectively. Table 3 (Appendix), reports the distribution of the span-level labels across splits.

4 Span Annotation

In this section, we describe our annotation framework including the manual annotation steps used for dataset construction, and the use of GPT-4 for different annotation roles. Figure 2 (Appendix) illustrates this framework.

4.1 Manual Annotation

The manual annotation process went through in two phases (Dimitrov et al., 2021a). For a given text $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ and a label (propaganda technique) space $\mathcal{Y} = \{y_1, y_2, \ldots, y_o\}$, each annotator A_i provides a set of spans S_{A_i} and each span is represented as $s_{A_i,y_j,k}$, where k is the index of the span for the i-th annotator and y_j is the label. Note that k can range from 1 to the total number of spans identified by annotator A_i, and this total can

[3] A submitted paper on this is currently under review.

be different for each annotator. Given this representation, for the i^{th} annotator the set of spans is defined as $S_{A_i} = \{s_{A_i,y_j,1}, s_{A_i,y_j,2}, \ldots, s_{A_i,y_j,m_i}\}$ where m_i is the total number of spans identified by annotator A_i and y_j represents any label from the label space, where j can vary from 1 to o. We combine the spans of all annotators into list $\mathcal{S}_\mathcal{C}$ that goes through the consolidation phase to finalize the annotations by consolidators.

To denote the labels (techniques) in a paragraph (input text $\mathbf{x}$) annotated by an annotator A_i, we define the following formulation: $\mathbb{Y}_{A_i} = \bigcup_{j=1}^{p} A_{i,y_j}$ where $\mathbb{Y}_{A_i}$ represents the set of all labels $\{y_1, y_2, \ldots, y_p\}$ annotated by A_i, where p is the total number labels. $\mathbf{Y}$ represents the list of labels from all annotators for a paragraph.

4.2 Annotation with GPT-4

To formally define the problem, let us consider the model $\mathcal{M}$, text input $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, and label space $\mathcal{Y}$. The task of $\mathcal{M}$ is to identify the text span $\mathcal{S} = \{s_1, s_2, \ldots, s_{m_i}\}$ and an associated label for each span s_i, where $s_i = y \in \mathcal{Y}$. The model is conditioned using instruction $\mathcal{I}$, which describes both the task and the label space $\mathcal{Y}$. This conditioning can occur in two scenarios: with a few-shot approach, utilizing labeled examples $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_l$, or in a zero-shot context, where labeled examples are not provided. $\mathcal{D}_l$ represents the labeled dataset. We formulated three levels of difficulty for the propaganda span annotation task using GPT-4.

- **Instruction only (Annotator)**: In this setup, the model is only provided with an instruction $\mathcal{I}$ asking it to annotate the text $\mathbf{x}$ by identifying the propaganda techniques used in it, and then extracting the corresponding spans $\mathcal{S}$.
- **Span extractor (Selector)**: We offer additional information for annotation and frame it as a span extraction problem. The model is asked to select the techniques manifesting in text from the list $\mathbf{Y}$, and extract the matching text spans.
- **Annotation consolidator (Consolidator)**: This setup is the most resource rich, where the model is asked to act as a consolidator, given list $\mathcal{S}_\mathcal{C}$ as provided by annotators.

5 Experimental Setup

In this section, we describe the setup of the experiments and the evaluation approach followed to investigate the effectiveness of GPT-4 in playing different roles in the annotation process.

Dataset. For the experiments in this study, we used the training subset of the dataset (discussed in Section 3) including 6,002 annotated paragraphs. In particular, we consider *the annotations resulting from the consolidation phase as our gold standard labels in all experiments*.

Model. For different experimental setups, we used zero-shot learning using GPT-4 (32K, version gpt-4-0314, temperature=0) (OpenAI, 2023). We chose this LLM due to its accessibility and superior performance compared to other models (Ahuja et al., 2023; Abdelali et al., 2024b).

Instruction. We specifically designed a prompt for each setup (as summarized in Table 4 in Appendix B). In prompting, we specifically ask the model to return the response in JSON format, given a provided template as reported in Table 4, to simplify parsing the output for such complex task.

5.1 Evaluation

We take two approaches to evaluate the performance of GPT-4 for our tasks.

Standard System Evaluation. We computed a modified version of the F_1 measure (macro- and micro-averaged) that accounts for partial matching between the spans across the gold labels and the predictions (Alam et al., 2022c).

Inter-rater Agreement. We also evaluated the quality of GPT-4's annotations through the computation of inter-rater agreement between its annotations and the gold labels from the dataset. We specifically computed γ (Mathet et al., 2015; Mathet, 2017), a measure used in similar tasks (Da San Martino et al., 2019), which is designed for span/segment-level annotation tasks.

6 Results and Discussion

To address our research questions, we ran each of the annotation setup prompts (Table 4, Appendix B) over all 6,002 paragraphs in the training split. Table 1 shows the results of evaluating the post-processed model's outputs.

Role	Micro-F_1	Macro-F_1	Span (γ)
Annotator	0.050	0.045	0.247
Selector	0.137	0.144	0.477
Consolidator	0.671	0.570	0.609

Table 1: Performance of GPT-4 (with its different roles) in propaganda span annotation using standard evaluation measures and annotation agreement.

Role	Micro-F$_{1orig}$	Micro-F$_{1correct}$
Annotator	0.050	0.117
Selector	0.137	0.297
Consolidator	0.671	0.670

Table 2: Performance of GPT-4 with ($_{correct}$) and without ($_{orig}$) span indices correction.

As shown in Table 1, the more information provided to GPT-4 during annotation, the more improvement we observed in its performance. In an information rich setup with GPT-4 as a "consolidator", where we used all the span-level annotations from three annotators, it led to significantly strong model performance. However, it should be noted that the task of a consolidator is not limited to deciding which of the initial annotations are the most accurate. They also had the freedom to modify the annotations by updating the annotation span length or by changing the label for a given span. As for annotation agreement, we can also see that the agreement scores were higher, when more information was provided to GPT-4 in the consolidator role, than the set ups with less information.

Incorrect start and end indices. In addition to detecting propaganda techniques, the model was required to provide the text spans matching these techniques (in the "annotator" and "selector" roles). Since a span might occur multiple times in a paragraph, with different context and propagandistic technique, the model should also specify the start and end indices of these spans. We observed that although GPT-4 can correctly provide labels and extract associated text spans, it frequently generated indices not matching the corresponding spans in a paragraph. This lead to mismatch between the start and end indices of spans as compared to gold labels (As Figure 1 (Appendix) shows). To overcome this problem, we a apply a post-processing step by assigning for each predicted span, the start and end indices of its first occurrence in a paragraph. Table 2 reports the performance of GPT-4 following this correction. It reveals the severity of inaccurate span positions prediction. With the first two roles of the model, we observe the performance increasing by a factor of two with the applied correction. Interestingly, in its third role, as a consolidator, this problem did not manifest, as the model was only selecting annotations, including span and indices, from the list $\mathcal{S}_C$ of all annotations.

Agreement with consolidators. We delve deeper into the quality of the model's annotations by comparing its agreement with consolidators (*after start indices correction*) to the agreement of the initial annotators with the consolidators. The dataset has an average agreement across annotators of γ = 0.531. For GPT-4, we observe a higher agreement score of 0.594 (as a selector) and 0.730 (as a consolidator). *These values demonstrate that GPT-4 achieves comparable or better agreement with the expert consolidators as compared to less experienced human annotators. Moreover, it shows that the model is learning from the given initial annotations to produce improved annotations, closer to the consolidator's performance.*

Per technique performance. We looked at the top per-technique agreement level (γ) of the model's labels versus gold labels (Table 5 in Appendix). Over all its roles, the model showed high agreement with expert annotators (consolidators) for three techniques: Doubt, Appeal to Hypocrisy and Loaded Language. The agreement was above 0.8 for at least 5 techniques.

Annotation cost. Manual annotation for the task is generally costly. The manual annotation of 6K paragraphs costed $3,600 for the team of 5 annotators.[4] As for using GPT-4, it resulted in reducing cost[5] by 96%, 36%, and 36% when acting as an annotator, selector, and consolidator, respectively.

7 Conclusions

In this study, we first investigate GPT-4's ability to play different roles in detecting propagandistic spans and annotating them in Arabic news paragraphs. We investigate if GPT-4 can be used as an annotator when provided with sets of information of varied richness, which represents an increased cost in hiring human annotators. Our experimental results suggest that providing more information significantly improves the model's annotation performance and agreement with human expert consolidators. The study also reveals the great potential of the model to replace consolidators, for some propaganda techniques. We offers an in-depth analysis of the model's performance across various annotation stages, facilitating a more informed adoption of this annotation approach. Future research will explore additional models and learning setups.

[4]The amount paid for annotators is very adequate given their country of residence.

[5]GPT-4 usage cost is computed by number of tokens per message (prompt and paragraph) sent to the model, in addition to initial annotations by 3 annotators for the last two roles.

8 Limitations

The current version of our work focuses on the analysis and evaluation of GPT-4 specifically limited to Arabic. For this study, we chose to use Arabic dataset because of the availability of annotated labels from multiple annotators, which are often difficult to obtain. We have evaluated only a closed large language model, as it is currently the most effective model for a large variety of NLP tasks and languages, as reported in a myriad of studies. Moreover, our experiments with large and effective open Arabic models for the task revealed that they are unable to understand the task.

Ethics and Broader Impact

We do not foresee any ethical issues in this study. We utilized an in-house dataset consisting of paragraphs curated from various news articles. Our analysis will contribute to the future development of datasets and resources in a cost-effective manner. Human annotators identity will not be shared and cannot be inferred from the annotations we plan to release. We would like to warn users to carefully use the annotations that we plan to release. It misuse (e.g., using them to generate similar content) may lead to a potential risk.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024a. LAraBench: Benchmarking Arabic AI with Large Language Models.

Ahmed Abdelali, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024b. LAraBench: Benchmarking Arabic AI with Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Malta. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022a. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Giovanni Da San Martino, Preslav Nakov, et al. 2022b. Overview of the {WANLP} 2022 shared task on propaganda detection in {A} rabic. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 108–118. Association for Computational Linguistics.

Firoj Alam, Hamdy Mubarak, Wajdi Zaghouani, Preslav Nakov, and Giovanni Da San Martino. 2022c. Overview of the WANLP 2022 shared task on propaganda detection in Arabic. In *Proceedings of the Seventh Arabic Natural Language Processing Workshop*, WANLP '22, Abu Dhabi, UAE.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks. *arXiv preprint arXiv:2307.02179*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management*, 56(5):1849–1864.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, EMNLP-IJCNLP 2019, Hong Kong, China.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021a. Detecting propaganda techniques in memes. *arXiv preprint arXiv:2109.08013*.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021b.

SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Maram Hasanain, Firoj Alam, Hamdy Mubarak, Samir Abdaljalil, Wajdi Zaghouani, Preslav Nakov, Giovanni Da San Martino, and Abed Freihat. 2023. ArAIEval shared task: Persuasion techniques and disinformation detection in Arabic text. In *Proceedings of ArabicNLP 2023*, pages 483–493, Singapore (Hybrid). Association for Computational Linguistics.

Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. 2023. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2020. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI '20, pages 4826–4832.

Yann Mathet. 2017. The agreement measure γ_{cat} a complement to γ focused on categorization of a continuum. *Computational Linguistics*, 43(3):661–681.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, 41(3):437–479.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI.

Andrew Perrin. 2015. Social media usage. *Pew research center*, pages 52–68.

Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, Jason Gonin, Camelia Ignat, Bonka Kotseva, Eleonora Mantica, Lorena Marcaletti, Enrico Rossi, Alessio Spadaro, Marco Verile, Giovanni Da San Martino, Firoj Alam, and Preslav Nakov. 2023a. News categorization, framing and persuasion techniques: Annotation guidelines. Technical report, European Commission Joint Research Centre, Ispra (Italy).

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023b. 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.

Piotr Przybyła and Konrad Kaczyński. 2023. Where does it end? long named entity recognition for propaganda detection and beyond. In *Proceedings of the International Conference of the Spanish Society for Natural Language Processing*.

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. Association for Computational Linguistics.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, ECML-PKDD '11, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI '22, pages 5597–5606, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large language models for propaganda detection. *arXiv 2310.06422*.

Nicolas Stefanovitch and Jakub Piskorski. 2023. Holistic inter-annotator agreement and corpus coherence estimation in a large-scale multilingual annotation campaign. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 71–86.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2023. Large language models can accurately predict searcher preferences. *arXiv preprint arXiv:2309.10621*.

Appendix

A Annotation Details

The annotation process of the dataset consisted of two phases: *(i)* in phase 1, three **annotators** individually annotated each paragraph, and *(ii)* in phase 2, two expert annotators revised and finalized the annotations (each annotator in this phase is referred to as a **consolidator**). To facilitate the annotation process, a platform was developed, and a comprehensive annotation guideline in the native language (Arabic) was provided to annotators and made available through out the process. Additionally, several training iterations were conducted before beginning the annotation task.

The two-phase annotation paradigm was designed to effectively approach the annotation task at hand, following extensive pilot studies on other paradigms, and quality assurance steps. Moreover, it is in essence similar to those followed in relevant studies (Piskorski et al., 2023a).

The reported annotation agreement for span-level annotation is $\gamma = 0.546$. This γ agreement metric is specifically designed for span/segment-level annotation tasks, taking into account the span boundaries (i.e., start and end) and their labels (Mathet et al., 2015; Mathet, 2017).

Table 3 provides the distributions of labels across the three dataset splits.

Technique	Train	Dev	Test
Appeal_to_Authority	192	22	42
Appeal_to_Fear-Prejudice	93	11	21
Appeal_to_Hypocrisy	82	9	17
Appeal_to_Popularity	44	4	8
Appeal_to_Time	52	6	12
Appeal_to_Values	38	5	9
Causal_Oversimplification	289	33	67
Consequential_Oversimplification	81	10	19
Conversation_Killer	53	6	13
Doubt	227	27	49
Exaggeration-Minimisation	967	113	210
False_Dilemma-No_Choice	60	6	13
Flag_Waving	174	22	41
Guilt_by_Association	22	2	5
Loaded_Language	7,862	856	1670
Name_Calling-Labeling	1,526	158	328
no_technique	2,225	247	494
Obfuscation-Vagueness-Confusion	562	62	132
Questioning_the_Reputation	587	58	131
Red_Herring	38	4	8
Repetition	123	13	30
Slogans	101	19	24
Straw_Man	19	2	4
Whataboutism	20	4	4
Total	**15,437**	**1,699**	**3,351**

Table 3: Distribution of the techniques in different data splits at the span level.

B Prompts

Table 4 lists the exact prompts used to invoke GPT-4 to act in its three different roles of interest in this work. During some pilot studies over the development subset, we have experimented with a variety of prompts for each of the roles before identifying the prompts we eventually used as they had the best performance in our pilot studies. We also note that model generally performed really well in responding with the required JSON format of output.

C Extended Results

Per technique performance. Our next research question is: which propaganda techniques can GPT-4 annotate best? We looked at the top five per-technique agreement levels (γ) of the model's labels versus gold labels (Table 5). Over all its roles, the model showed high agreement with expert annotators (consolidators) for three techniques: Doubt, Appeal to Hypocrisy and Loaded Language.

It is interesting to see that GPT-4 was highly effective in annotation of the "Doubt" technique, which contradicts with a recent ranking of annotation difficulty of the same taxonomy, derived from humans' performance, in the same task across a multilingual dataset (Stefanovitch and Piskorski, 2023). However, its strong performance with the other two techniques is inline with the aforementioned ranking. The model's ability to annotate "Loaded Language" is particularly useful, as it is the most prevalent technique in the dataset, appearing 7.9K times in the training split under investigation. Replacing human consolidators by GPT-4 to annotate for that technique can save tremendous time and cost. We believe these agreement levels give further evidence of the strong potential of employing GPT-4 as a propaganda span annotator, at least for some techniques. This analysis also provides data needed to inform decisions on which stages of annotation we can inject LLMs like GPT-4.

Setup	Prompt
Annotator	**Instruction** ($\mathcal{I}$)**:** Label the "Paragraph" by the following propaganda techniques: [techniques list]. Answer exactly and only by returning a list of the matching labels from the aforementioned techniques and specify the start position and end position of the text span matching each technique. Use this template {"technique": , "text": , "start": , "end": } **Paragraph**: {. . . } **Response**:
Selector	**Instruction** ($\mathcal{I}$)**:** Given the following "Paragraph" and "Annotations" showing propaganda techniques potentially in it. Choose the techniques you are most confident appeared in Paragraph from all Annotations and return a Response. Answer exactly and only by returning a list of the matching labels and specify the start position and end position of the text span matching each technique. Use this template Use this template {"technique": , "text": , "start": , "end": } **Paragraph**: {. . . } **Annotations**: $\mathbf{Y}$ **Response**:
Consolidator	**Instruction** ($\mathcal{I}$)**:** Given the following "Paragraph" and "Annotations" showing propaganda techniques potentially in it, and excerpt from the Paragraph where a technique is found. Choose the techniques you are most confident appeared in Paragraph from all Annotations and return a Response. Answer exactly and only by returning a list of the matching annotations. **Paragraph**: {. . . } **Annotations**: $\mathcal{S}_\mathcal{C}$ **Response**:

Table 4: Different prompts used to instruct GPT-4 to annotate input paragraphs by propaganda techniques and spans.

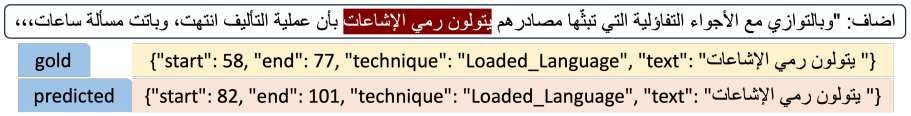
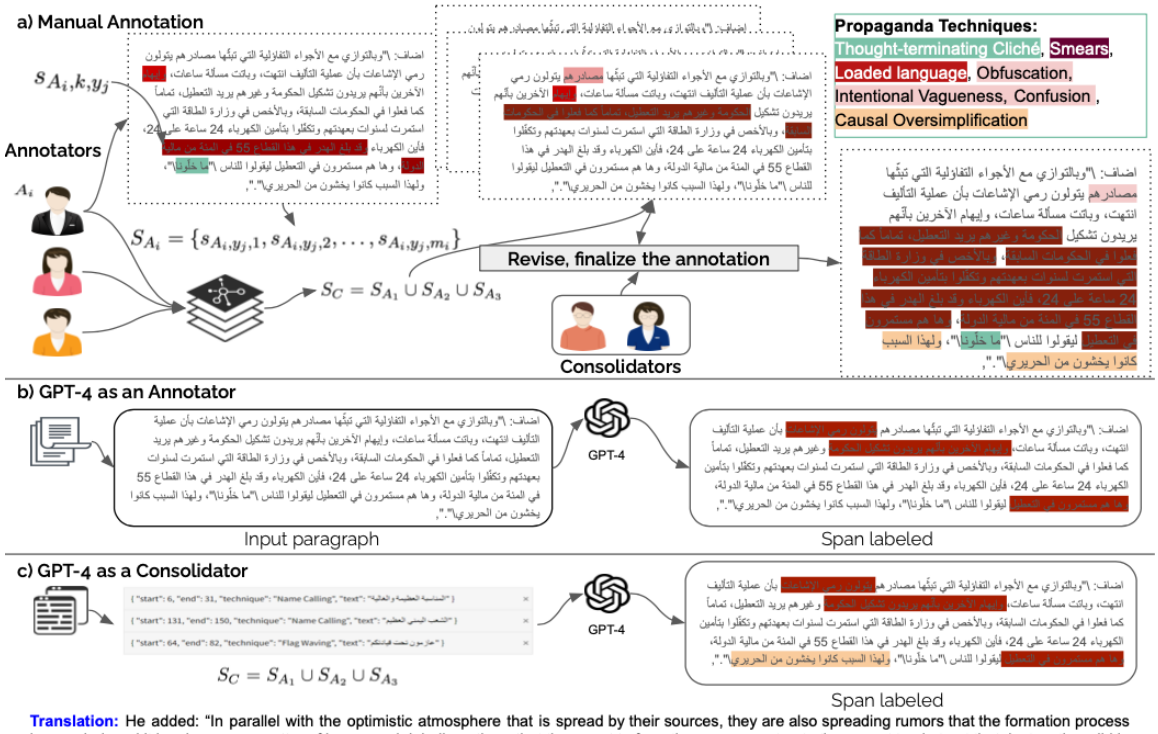


Figure 1: Example of wrongly generated span indices by GPT-4.

Technique	Annotator
Causal Oversimplification	0.889
Consequential Oversimplification	0.835
Doubt	0.815
Obfuscation /Vagueness /Confusion	0.791
Appeal to Hypocrisy	0.746
	Selector
Doubt	0.802
Flag Waving	0.705
Appeal to Hypocrisy	0.660
Loaded Language	0.654
Slogans	0.642
	Consolidator
False Dilemma /No Choice	0.872
Loaded Language	0.774
Straw Man	0.697
Doubt	0.695
Name Calling /Labeling	0.680

Table 5: Agreement level (measured by γ) between GPT-4 and gold labels for top five techniques per role, with ($_{correct}$) span indices correction. Underlines are those techniques appearing in at least two annotation roles.



Translation: He added: "In parallel with the optimistic atmosphere that is spread by their sources, they are also spreading rumors that the formation process has ended, and it has become a matter of hours, and deluding others that they want to form the government nut others want to obstruct that, just as they did in previous governments, especially in the Ministry of Energy, which continued for years under their charge and they were responsible for ensuring that electricity was available 24 hours in 24 hours , so where is the electricity, when the waste of money in this sector has reached 55 percent of the state's finances, but here they are continuing to obstruct to tell the people that ``they didn't let us" and for this reason they were afraid of Hariri."

Figure 2: Existing span-level annotation process requiring human annotators and expert consolidators, while our proposed solution uses GPT-4 to support annotation and consolidation.