Council of LLMs: Evaluating Capability of Large Language Models to Annotate Propaganda

Vivek Sharma[1], Shweta Jain[1,2], Mohammad Mahdi Shokri[1], Sarah Ita Levitan[3,1], Elena Filatova[4,1]
[1]The Graduate Center, CUNY, [2]John Jay College of Criminal Justice, CUNY,
[3]Hunter College, CUNY, [4]New York City College of Technology, CUNY,

Abstract

Data annotation is essential for supervised natural language processing tasks but remains labor-intensive and expensive. Large language models (LLMs) have emerged as promising alternatives, capable of generating high-quality annotations either autonomously or in collaboration with human annotators. However their use in autonomous annotations is often questioned for their ethical take on subjective matters. This study investigates the effectiveness of LLMs in an autonomous, and hybrid annotation setups in propaganda detection. We evaluate GPT and open-source models on two datasets from different domains, namely, Propaganda Techniques Corpus (PTC) for news articles and the Journalist Media Bias on X (JMBX) for social media. Our results show that LLMs, in general, exhibit high recall but lower precision in detecting propaganda, often over-predicting persuasive content. Multi-annotator setups did not outperform the best models in single-annotator setting although it helped reasoning models boost their performance. Hybrid annotation, combining LLMs and human input, achieved the highest overall accuracy than LLM-only settings. We further analyze misclassifications and found that LLM have higher sensitivity towards certain propaganda techniques like loaded language, name calling, and doubt. Finally, using error typology analysis, we explore the reasoning provided on misclassifications by the LLM. Our result shows that although some studies report LLM outperforming manual annotations and it could prove useful in hybrid annotation, its incorporation in the human annotation pipeline must be implemented with caution.

1 Introduction

Data annotation is a foundational task in natural language processing (NLP), providing the ground truth needed to train and evaluate supervised models. Traditionally, this task is performed by subject matter experts (SMEs) who follow detailed guidelines to ensure consistency and accuracy. Due to cost and time consideration the annotation process is often scaled using crowdsourcing platforms like Amazon Mechanical Turk and Prolific Fort et al., Sabou et al.. These labels are typically filtered and validated to produce high-quality "gold" annotations to train NLP models.

Despite its importance, manual annotation remains slow, expensive, and susceptible to annotator fatigue, especially with large datasets. This can lead to inconsistencies that degrade model performance. In response, recent work has explored using large language models (LLMs), such as GPT-4 and GPT-4o, for annotation. These models have shown promise in generating high-quality labels in both hybrid settings, where human annotators review LLM-generated suggestions and autonomous settings, where models label data without human intervention.

Hybrid annotation seeks to reduce human workload while preserving annotation quality. In contrast, fully automated annotation offers scalability but raises concerns about bias, hallucination, and the propagation of model-generated errors. While LLMs continue to improve in performance and alignment, questions remain about their reliability and accountability in annotation workflows.

Though LLMs have been evaluated for tasks like classification, summarization, and question answering, their use as a multi-annotator annotator in subjective tasks like propaganda detection remains underexplored. Moreover, there is a lack of in-depth exploration in their reasoning which can help us understand the black-box nature of their analysis. This paper addresses the gap in the context of propaganda detection, a domain where language is often subtle and context-dependent. It is easy for annotators' biases to impact labels in the data. In contrast, LLMs can be engineered to act neutrally and adhere to guidelines more strictly. We study the difference in annotation accuracy compared to

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 1–12

human-generated labels by measuring LLMs' capability in single-annotator, multi-annotator, and hybrid settings in an exhaustive comparison of the latest GPT Achiam et al. and open-source models. We evaluate LLMs using two datasets: the Propaganda Techniques Corpus (PTC) (Martino et al., 2020) for news articles and Journalist Media Bias on X (JMBX) (Sharma et al., 2025a) for social media content.

Through this study, we investigate the potential of LLMs to perform annotation without human supervision and evaluate the performance of multi-annotator collaboration among LLMs, both quantitatively and qualitatively. Specifically, we explore the following research questions:

RQ1: Does hybrid annotation(human-LLM) perform better in detecting propaganda than automated LLM annotations? To what extent does hybrid annotation (human–LLM) outperform fully automated LLM annotation in detecting propaganda?

RQ2: Do human–LLM disagreements cluster around specific propaganda techniques, and if so, which ones?

RQ3: Under what conditions, if any, does consolidation in a multi-annotator setting match or exceed single-annotator accuracy across text domains (news vs. social media)?

Our results underscore shortcomings of multi-annotator models and how consolidation performance in adjudication depends on individual annotators. Qualitatively, we find that while proprietary and open-source LLMs often detect valid propaganda patterns, they overgeneralize—especially in informal social media contexts. Their behavior is shaped by linguistic features as well as formatting and tone, which can cause systematic mislabeling without strong contextual grounding.

2 Related Work

LLM perform well on domain-specific annotation tasks, including software engineering (Ahmed et al., 2025). Beyond technical domains, Törnberg shows GPT models can surpass expert annotators and supervised classifiers in labeling political social media. These results suggest LLM annotations can closely match human work. However, several scholars warn of risks in social science annotation workflows. LLMs are often viewed as black boxes (Kristensen-McLachlan et al., 2023; Bender et al., 2021), making their output difficult to interpret and reproduce (Ollion et al., 2024). Moreover, their training data can embed and amplify existing biases, such as gender stereotypes (Lucy and Bamman, 2021). By contrast, human annotation follows guidelines and quality controls to mitigate such biases. Consequently, calls have been made for standards for LLM-based annotations to ensure reliability and transparency (Törnberg, 2024a,b).

Technical limitations constrain the efficacy of LLMs by being prone to hallucination (Lin and Zhang, 2025), over-confidence (Xiong et al., 2023), and when annotations are wrong, often produce explanations lacking coherence and sufficiency (Wang et al., 2024). Prompt design can improve annotation quality (Reiss, 2023), though others find only marginal gains (Rytting et al., 2023; Murugadoss et al., 2025). Societal risks persist: removing human annotators can erase interpretive insight and reinforce systemic bias (Das et al., 2024; Abid et al., 2021; Wang et al., 2023), leading to unfair outcomes and broader disparities (Dillion et al., 2023).

To balance risks and benefits, hybrid workflows let LLMs and humans collaborate across annotation stages. LLMs may pre-process data, propose labels, or validate human work. Lin and Zhang (2025) position the LLM either as the main annotator for large-scale tasks with humans verifying output or as a secondary assistant when dataset sizes are smaller. Several studies have experimented with different hybrid configurations. Heseltine and Clemm von Hohenberg (2024) use a three-phase pipeline where the LLM labels twice, discrepancies are reconciled, then a final model is trained. Hamilton et al. (2024) introduce RhetAnn, a web tool that aids annotators with LLM-generated explanations for rhetorical techniques. Wang et al. (2024) adds LLM explanations as supplementary guidance, accelerating work while keeping humans in the loop. Hasanain et al. (2023) shows LLMs can serve as consolidators, roles usually for experts, achieving strong propaganda span detection with detailed rationales. Kim et al. (2024) showed hybrid frameworks improves quality and address scalability and fatigue. Rouzegar and Makrehchi (2024) integrate LLM uncertainty into active learning, switching between human and machine annotators by confidence thresholds.

As hybrids mature, fully automated annotation is also examined. Hamilton et al. (2024) report GPT-4 matching human quality at one-tenth the cost with

minimal manual labels. Törnberg (2023) finds that ChatGPT outperforms crowdsourced annotators on tweets. Not all agree: Gu et al. (2025) show automated LLMs beat TF-IDF methods yet remain less reliable than experts. Golazizian et al. (2024) combine RoBERTa-Large and LLama-3, gaining 4% F1 on hate speech with 50% of the budget and 2% on moral sentiment with 25% of costs. Subramaniam et al. (2025) proposed a self-refining multi-annotator system that preserves diverse reasoning.

These studies show multi-annotator LLMs and hybrid annotations can handle varied tasks with mixed performance. However, none directly compare single-annotator and multi-annotator models on subjective tasks like propaganda detection. Such a comparison helps assess performance differences and reveal how LLMs behave in multi-annotator settings. Grasping classification nuances is necessary to interpret LLM reasoning, see where it diverges from human annotation, and analyze that process. This, in turn, helps prompt engineers refine prompts and adjust model parameters for subjective tasks. Our aim is not peak accuracy, but to explore LLM performance across scenarios and domains.

3 Experiment and Result

To evaluate LLM annotation in single-annotator, multi-annotator, and hybrid setups, we run experiments on two datasets: the Propaganda Techniques Corpus (PTC) Martino et al. (2020) and the Journalist Media Bias on X (JMBX) corpus Sharma et al. (2025b). PTC provides sentence-level binary labels for news articles, while JMBX contains annotated tweets from journalists' accounts associated with biased outlets on X (formerly Twitter). From each dataset, we sample 200 sentences with equal numbers of propaganda and non-propaganda instances. The propaganda subset follows the natural distribution of techniques in the original datasets. All models, both GPT and open-source LLMs, are evaluated on the same sentences for consistency and the reported performance is the average of three runs on each setting.

In standalone setting, each model independently performs binary classification (propaganda vs. non-propaganda) at the sentence level. Using a prompt that includes technique definitions by Martino et al. (2020) (see Appendix B), the model is asked to identify any technique(s) present. If a technique is returned, the sentence is labeled as propaganda; if none is returned, it is labeled as non-propaganda. In a multi-annotator setting, each model first operates as an independent annotator. In this stage, the model labels every sentence with one or more propaganda techniques and provides a brief justification for each decision. These initial annotations are generated in isolation, so no model has access to the labels or rationales produced by any other model.

In the second stage, each model is then anonymously presented with the full set of sentence-level labels and justifications produced in the first stage by all models, including its own. Using a consolidation prompt that includes the definitions of all propaganda techniques, the model is instructed to review these candidate labels and select a single final label together with its accompanying justification. Importantly, this consolidation step is framed as a re-annotation and not as an aggregation procedure

We replicate the same prompts and sentence sets across proprietary GPT models and open-source models on both PTC and JMBX to compare performance across models and datasets.

The performance of LLMs in various configuration is reported using standard metrics (Sokolova and Lapalme, 2009): F1-score, precision and recall generally used in the studies. This analysis is followed by qualitative analysis understanding understanding the reasoning on misclassifications.

3.1 Automated annotations by LLM

3.1.1 GPT models

First, we evaluate the annotation performance of three GPT models: GPT-4o (gpt-4o-2024-08-06), GPT-4.1 (gpt-4.1-2025-04-14), and GPT-o3 (o3-2025-04-16). These models are selected for their diverse capabilities. GPT-4.1 for its proficiency in complex reasoning, GPT-4o for its speed and balanced performance, and GPT-o3 for its specialized reasoning strengths simulating varied annotator expertise.[1]

Table 1 and Table 2 report performance on PTC and JMBX datasets, respectively.

[1] https://platform.openai.com/docs/models

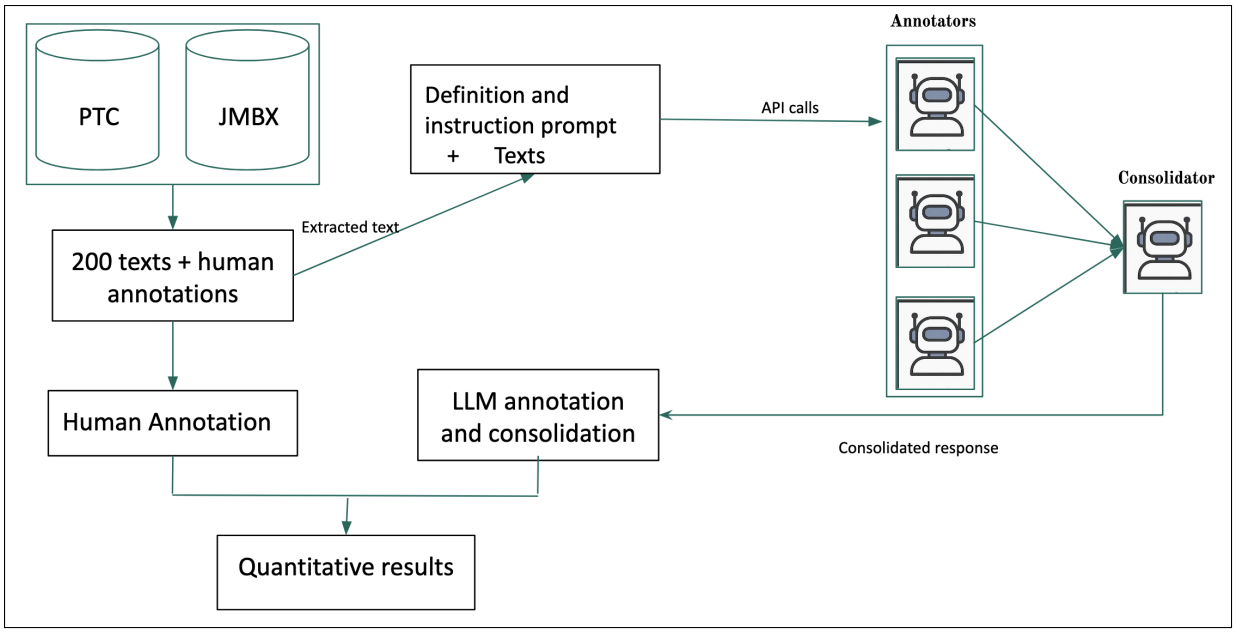

Figure 1: Experiment to assess LLMs performance in annotating propaganda in an automated setting

Model	P/R/F	Consolidation P/R/F
GPT 4.1	**0.77/0.74/0.74**	0.78/0.71/0.70
GPT 4o	0.79/0.74/0.73	0.78/0.70/0.68
GPT o3	0.78/0.70/0.68	0.78/0.71/0.69

Table 1: Performance of GPT models in annotating PTC dataset as a standalone system and as consolidator in a multi-annotator setting

Model	P/R/F	Consolidation P/R/F
GPT 4.1	**0.77/0.77/0.76**	0.74/0.72/0.71
GPT 4o	**0.77/0.77/0.76**	0.74/0.71/0.71
GPT o3	0.73/0.71/0.70	0.77/0.74/0.73

Table 2: Performance of GPT models in classifying JMBX dataset as a standalone system and as consolidator in a multi-annotator setting

Our analysis shows that the flagship GPT models performed well in a single-annotator setup on both datasets compared to the reasoning models. However, their performance declined in the multi-annotator setup, while the reasoning model(o3) improved. This highlights that performance in a multi-annotator setup depends on annotator quality: when annotators perform poorly, they reduce the consolidator's effectiveness, even when the consolidator is instructed not to aggregate results but to decide after careful analysis.

3.1.2 Open-source LLMs

We repeat the same experiments using open-source LLMs: Qwen3-32B, QwQ-32B, and DeepSeek-R1 Llama 3.3 70B. The performance in standalone and consolidator configurations is shown in Table 3 (PTC) and Table 4 (JMBX). The result shows a similar pattern to the proprietary models. The reasoning model here, QwQ, performed well in the multi-annotator setting, whereas the other two models performed better in the single-annotator setting and the performance dipped due to a poor performance of the reasoning model. This shows that consolidation is only as good as the individual annotations in a multi-annotator setting.

Model	P/R/F	Consolidation P/R/F
Qwen3-32B	0.78/0.76/0.75	**0.79/0.73/0.72**
QwQ-32B	0.79/0.70/0.68	0.79/0.72/0.71
DeepSeek R1*	**0.80/0.76/0.75**	0.79/0.73/0.71

Table 3: Performance of open-source LLMs in annotating PTC dataset as a standalone system and as consolidator.*DeepSeek-R1-Distill-Llama-70B

Model	P/R/F	Consolidation P/R/F
Qwen3-32B	**0.70/0.70/0.70**	**0.71/0.69/0.69**
QwQ-32B	0.70/0.65/0.63	0.72/0.69/0.67
DeepSeek R1*	0.70/0.69/0.68	0.68/0.66/0.65

Table 4: Performance of open-source LLMs in annotating JMBX dataset as a standalone system and as consolidator.*DeepSeek-R1-Distill-Llama-70B

Overall, open-source LLMs performed better on news articles than on social media texts.

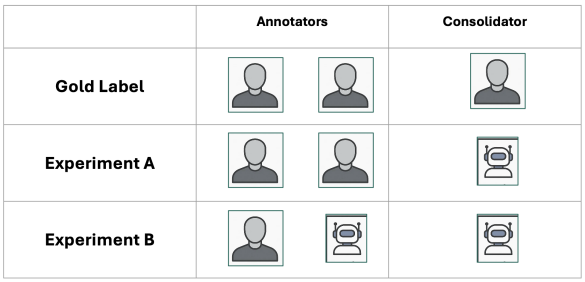

Figure 2: Hybrid Annotation experiments

Among these, DeepSeek-R1 showed consistently strong performance across tasks. Interestingly, the reasoning-focused models (GPT-o3 and QwQ-32B) had the weakest performance when used alone; however, their scores improved in multi-annotator setups, likely benefiting from the additional contextual input provided by other better-performing models.

3.2 Hybrid annotation

To complete our evaluation, we conducted Experiment 3, involving a hybrid setup with both human and LLM collaborators. Because the JMBX dataset includes annotations from two human annotators and one human consolidator, it enabled us to simulate and compare mixed human–AI annotation workflows.

In the first hybrid configuration, annotations by human annotators were passed to LLM models for final consolidation. The results are shown in Table 5, revealing that hybrid setups outperformed fully automated configurations.

Model	Consolidation P/R/F	Class-wise P/R P(prop), R(prop), P(non-prop), R(non-prop)
GPT 4.1	0.79/0.79/0.78	0.76/0.83/0.81/0.74
GPT 4o	**0.85/0.84/0.84**	0.86/0.83/0.83/0.86
GPT o3	0.78/0.76/0.75	0.70/0.90/0.86/0.61

Table 5: Performance of hybrid setup (experiment A) with GPT models as consolidators on the JMBX dataset. P/R/F are weighted-averaged. The last column lists class-wise P/R as *(prop, non-prop)*.

Model	Consolidation P/R/F	Class-wise P/R P(prop), R(prop), P(non-prop), R(non-prop)
Qwen3-32B	**0.76/0.76/0.75**	0.74/0.79/0.77/0.72
QwQ-32B	0.75/0.69/0.66	0.62/0.94/0.88/0.43
DeepSeek R1*	0.74/0.70/0.69	0.65/0.89/0.83/0.52

Table 6: Performance of hybrid setup (experiment A) with open-source models as consolidators on the JMBX dataset.*DeepSeek-R1-Distill-Llama-70B P/R/F are weighted-averaged across classes. The last column lists class-wise P/R as *(prop, non-prop)*.

In a follow-up scenario, we replaced annotations by one human annotator and human consolidator by LLMs, keeping only one human annotation in the loop. GPT-4o acted as the consolidator and GPT-4.1 as the other annotator, while, in a parallel open-source configuration, DeepSeek R1 acted as the consolidator and Qwen3 as the annotator. GPT-4o's performance with precision/recall/F1 = 0.79/0.79/0.79, with class-specific scores of 0.70/0.90 for propaganda and 0.86/0.61 for non-propaganda. As expected, the performance dropped slightly compared to the previous hybrid setup, but it still exceeded the GPT-4o model's performance in its standalone and consolidation roles (see Table 2). A similar trend was observed with the open-source models, where precision/recall/F1 was 0.75/0.74/0.74, with class-specific scores of 0.74/0.76 for propaganda and 0.75/0.73 for non-propaganda.

4 Performance Analysis

To better understand model behavior and failure points, we analyze misclassifications across both the propaganda and non-propaganda classes. Specifically, we examine class-wise precision and recall to identify which classes the models over-predict or struggle to detect accurately. These detailed metrics are shown in Appendix C Tables 7 through 10.

A consistent pattern across models is low precision and high recall for the propaganda class. This indicates that models tend to over-predict propaganda, often labeling neutral sentences as propagandistic. Among the GPT models, performance is generally stronger on social media text than on news articles. Notably, GPT-4o performs particularly well on JMBX with balanced results: 0.78 precision and 0.73 recall for propaganda, and 0.75 precision and 0.80 recall for non-propaganda. This suggests GPT-4o may be more attuned to the infor-

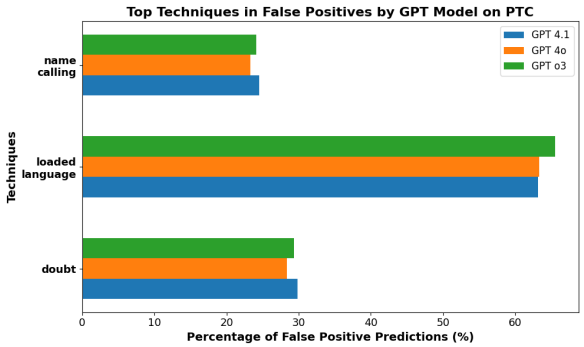

Figure 3: Top Propaganda Techniques in false positive by GPT models on PTC dataset

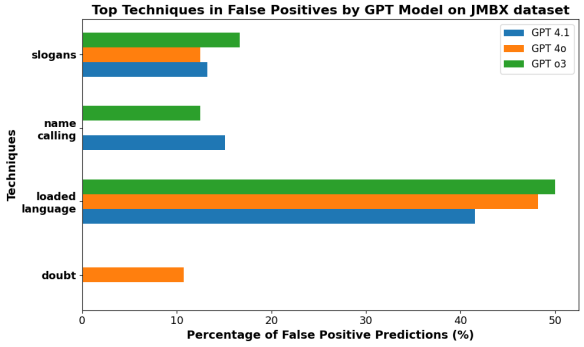

Figure 4: Top Propaganda Techniques in false positive by GPT models on JMBX dataset

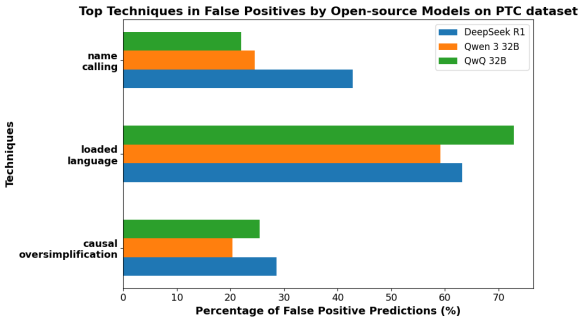

Figure 5: Top Propaganda Techniques in false positive by open source models on PTC dataset

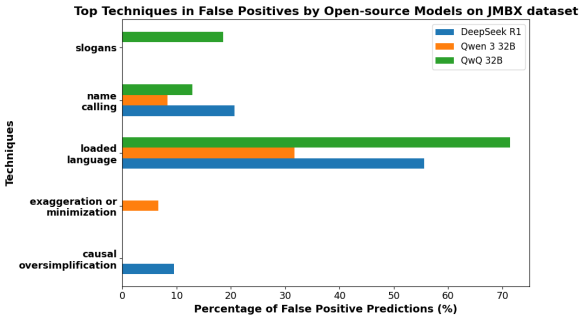

Figure 6: Top Propaganda Techniques in false positive by open source models on JMBX dataset

mal language styles common in social media.

Open-source LLMs show high recall but relatively low precision for propaganda, mirroring the aggressive classification pattern seen in GPTs. Qwen3-32B, however, stands out with the most balanced class-wise performance across both datasets. For PTC, it achieves a precision of 0.70 and recall of 0.90 for propaganda, while also maintaining relatively high scores for non-propaganda (0.86 precision, 0.61 recall). On the JMBX dataset, Qwen3-32B remains strong and balanced, outperforming other local models in non-propaganda recall.

Given the high false positive rate across models, we examine the propaganda techniques where human–LLM disagreement was greater. Although a full fine-grained analysis is beyond the scope of this work, we used technique labels assigned by the models in false-positive cases to inspect which rhetorical patterns were overused.

As shown in Figures 3 and 4, "Loaded Language" was the most common technique in misclassifications by GPT models. This suggests that GPTs are highly sensitive to emotionally charged or emphatic wording, even in non-propagandistic sentences. The second most common techniques were "Name Calling" and "Doubt," which may indicate a bias toward detecting adversarial or skeptical tones Martino et al.. On the JMBX dataset specifically, false positives frequently included "Slogans," likely triggered by tweet-formatting conventions such as all caps or punchy phrases.

Open-source LLMs demonstrated similar trends. As shown in Figures 5 and 6, "Loaded Language" again appeared as the dominant technique in false positives. However, these models also frequently mislabeled sentences with "Causal Oversimplification" and "Exaggeration or minimization" indicating error patterns that differ from those of the GPT models. These techniques often involve broad generalizations or emotionally overstated claims, which may be harder to contextualize without deeper sentence-level understanding.

5 Qualitative Analysis of human-LLM disagreements

Quantitative metrics are necessary but insufficient when interpretive ambiguity and reasoning quality matter: a model may diverge from a gold label yet offer a defensible rationale. Recent work by Belinkov and Glass (2019) calls for qualitative analyzes to expose such nuances and complement ag-

gregate scores. Error-typology analysis, in particular, groups mismatches into cohorts to reveal failure patterns, overreach, and reasonable disagreement, and prior studies urge looking beyond fluent outputs to the quality of underlying reasoning (Narang et al., 2020).

In this section, we run a structured qualitative analysis of LLM–human disagreements, focusing on prevalent false positives where GPT predictions diverge from human gold labels across both datasets. Our approach draws on behavioral testing and cohort-based diagnostics (e.g., GLUE diagnostics, NLI stress tests), where naming reproducible error types supports targeted evaluation and debugging.

Methodologically, we treat disagreement as signal, consistent with evidence that some conflicts reflect genuine ambiguity rather than annotator error (Pavlick and Kwiatkowski, 2019). We group mismatches into three categories by the interpretive value of the LLM's explanation: (i) Defensible Divergences (reasonable alternate readings), (ii) Overreaching Reasoning (technically plausible but contextually weak), and (iii) Unconvincing/Unsupported Reasoning (clear failures). This reason-aware framing complements standard metrics and helps separate true model errors from interpretive disagreements.

5.1 Defensible Divergence

In this case, the LLM prediction does not match the gold label, but the reasoning depicts rhetorical awareness and interpretive depth that makes the label plausible. These are not errors in the conventional sense, but alternative valid readings of rhetorical cues. This category highlights a key limitation of rigid gold labels and bias of human subjects in subjective tasks.

Text:*"What could possibly go wrong at Norwich?*

- **LLM annotation:** Propaganda(Doubt)
- **Justification**: The sentence 'what could possibly go wrong at Norwich?' is utilizing the technique of doubt by questioning the credibility or reliability of the situation or event occurring at Norwich. It implies skepticism about the outcome without providing specific details or evidence, leading the audience to question or doubt what might occur there.

The above example suggest that strict disagreement with gold labels may underestimate the interpretive capacity of LLMs, especially when the model is attuned to subtle rhetorical stimulation.

5.2 Reasonable but Overreaching

In the following example, the LLM's reasoning follows the formal definition of a propaganda technique, but its application feels forced or contextually inappropriate. These mismatches reflect overgeneralization where the model applies technique labels to rhetorical elements that are technically present but not persuasive or manipulative in intent.

Text:*"As a firearms expert, he questioned the lack of flashes coming from the hotel windows where the shooter was supposedly firing from."*

- **LLM annotation:** Propaganda(Appeal to Authority, Doubt)
- **Human annotation:** Non-propaganda
- **Justification**: The sentence uses 'as a firearms expert' to give credibility to the speaker's claim; it also includes 'questioned the lack of flashes', which fits the doubt technique.

The example highlight how surface-level match to technique definition can mislead models when they are not grounded in pragmatic or discourse-aware understanding.

5.3 Unconvincing or Unsupported Reasoning

In the final category, LLM neither aligns with the gold label nor offers a compelling justification for its choice. The model's explanations rely heavily on partial lexical cues or structural repetition, without demonstrating real understanding of rhetorical function or context. Here's an example of model failure.

Text:*"Completing this poll grants you access to Freedom Outpost updates free of charge."*

- **LLM annotation:** Propaganda(Loaded Language)
- **Justification**: The term 'freedom' in 'Freedom Outpost' carries a strong positive emotional connotation, appealing to readers' favorable feelings toward the concept of freedom.

This qualitative analysis reveals that not all mismatches are same. While some represent genuine failures, others demonstrate defensible divergence

or overgeneralization, both of which carry implications for how we should evaluate LLM performance and question over-reliance on human annotation as gold label for subjective task. These divergences can prove its usefulness to extend human-insights or question human's decision in their annotation. Through this we aim to show the nuance in model's performance and show importance of holistic perception in such task.

6 Conclusion

This study investigates LLMs for propaganda annotation in single annotator, multi-annotator, and hybrid configurations. We test multiple proprietary GPT models and open-source models (e.g., Qwen, DeepSeek R1) on news and social media datasets to evaluate annotation capability and identify failure points.

LLM performance falls short of human accuracy. Human–LLM hybrids consolidated by GPT-4o come closest to experts, followed by single-annotator LLM. Multi-annotator LLMs perform worst, with outcomes tied to the accuracy of the consolidated individual annotations.

Most models show low recall for the non-propaganda class, yielding many false positives. To probe these errors, we analyzed techniques assigned in false positives. "Loaded Language" was most overused, followed by "Name Calling" and "Doubt"; open source models sometimes mapped non-propagandistic content to "Causal Oversimplification." This indicates high sensitivity to certain techniques, so neutral content is often labeled accordingly.

Using an error typology, we categorized human–LLM disagreements into three groups based on LLMs' stated reasoning. These categories show misclassifications are not equal, informing debates on LLMs in annotation workflows and highlighting both potential and limitations.

7 Limitations

While our experiments offer useful insights into LLM capabilities for propaganda detection in standalone and multi-annotator settings, several limitations remain. Our analysis centers on binary classification. Although we include exploratory technique-level analysis in our qualitative evaluations, we do not perform full multi-label classification. Consequently, the models' ability to identify multiple propaganda spans within a single sentence is underexplored. Additionally, the number of sentences included in our evaluation is relatively small due to budgetary and token-usage constraints, which limited the scale of our experimental setup. These limitations suggest several future directions: larger-scale evaluations, fine-grained multi-label annotation, iterative multi-annotator collaboration, and human-in-the-loop verification.

8 Ethical Considerations

This study is limited to propaganda detection using fixed definitions and task-specific prompts. A key concern is bias in LLM training data, which may influence annotation outcomes despite objective definitions. While we did not evaluate bias directly, its impact on reliability, especially in politically or culturally sensitive contexts, warrants further investigation.

We also caution against unsupervised pipelines that use LLM-generated annotations as training data. LLMs can hallucinate or misclassify ambiguous inputs, which is problematic in high-stakes domains. Mislabeling satire, dissent, or stylistic variation as propaganda can have ethical consequences in real-world deployments. Finally, while LLMs can support or automate annotation, human oversight remains essential, especially for consolidating judgments, handling edge cases, and ensuring transparency and accountability. LLMs can benefit further from human–LLM collaboration via iterative feedback.

9 Acknowledgment

This work used computational resources provided by the National Artificial Intelligence Research Resource (NAIRR) Pilot under Allocation NAIRR240383.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Toufique Ahmed, Premkumar Devanbu, Christoph Treude, and Michael Pradel. 2025. Can LLMs replace manual annotation of software engineering ar-

tifacts? In *Proceedings of the 22nd IEEE/ACM International Conference on Mining Software Repositories (MSR)*, pages 526–538. IEEE.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Amit Das, Zheng Zhang, Najib Hasan, Souvika Sarkar, Fatemeh Jamshidi, Tathagata Bhattacharya, Mostafa Rahgouy, Nilanjana Raychawdhary, Dongji Feng, Vinija Jain, and 1 others. 2024. Investigating annotator bias in large language models for hate speech detection. In *Neurips Safe Generative AI Workshop 2024*.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Karën Fort, Gilles Adda, and Kevin Bretonnel Cohen. 2011. Amazon mechanical turk: Gold mine or coal mine? *Computational Linguistics*, 37(2):413–420.

Preni Golazizian, Alireza Salkhordeh Ziabari, Ali Omrani, and Morteza Dehghani. 2024. Cost-efficient subjective task annotation and modeling through few-shot annotator adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3474–3491.

Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. 2025. Large language models are effective human annotation assistants, but not good independent annotators. *arXiv preprint arXiv:2503.06778*.

Kyle Hamilton, Luca Longo, and Bojan Bozic. 2024. Gpt assisted annotation of rhetorical and linguistic features for interpretable propaganda technique detection in news text. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1431–1440.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2023. Large language models for propaganda span annotation. *arXiv preprint arXiv:2311.09812*.

Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239.

Hannah Kim, Kushan Mitra, Rafael Li Chen, Sajjadur Rahman, and Dan Zhang. 2024. Meganno+: A human-llm collaborative annotation system. *arXiv preprint arXiv:2402.18050*.

Ross Deans Kristensen-McLachlan, Miceal Canavan, Márton Kardos, Mia Jacobsen, and Lene Aarøe. 2023. Chatbots are not reliable text annotators. *arXiv preprint arXiv:2311.05769*.

Hao Lin and Yongjun Zhang. 2025. The risks of using large language models for text annotation in social science research. *arXiv preprint arXiv:2503.22040*.

Li Lucy and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the third workshop on narrative understanding*, pages 48–55.

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*.

Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2025. Evaluating the evaluator: Measuring llms' adherence to task evaluation instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19589–19597.

Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*.

Étienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2024. The dangers of using proprietary llms for research. *Nature Machine Intelligence*, 6(1):4–5.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Michael V Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Hamidreza Rouzegar and Masoud Makrehchi. 2024. Enhancing text classification through llm-driven active learning and human annotation. *arXiv preprint arXiv:2406.12114*.

Christopher Michael Rytting, Taylor Sorensen, Lisa Argyle, Ethan Busby, Nancy Fulda, Joshua Gubler, and David Wingate. 2023. Towards coding social science datasets with language models. *arXiv preprint arXiv:2306.02177*.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866. Citeseer.

Vivek Sharma, Mohammad Shokri, Sarah Levitan, Elena Filatova, and Shweta Jain. 2025a. Analysis of propaganda in tweets from politically biased sources.

The International FLAIRS Conference Proceedings, 38.

Vivek Sharma, Mohammad Mahdi Shokri, Sarah Ita Levitan, Elena Filatova, and Shweta Jain. 2025b. Analysis of propaganda in tweets from politically biased sources.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Vighnesh Subramaniam, Yilun Du, Joshua B Tenenbaum, Antonio Torralba, Shuang Li, and Igor Mordatch. 2025. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*.

Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588*.

Petter Törnberg. 2024a. Best practices for text annotation with large language models. *arXiv preprint arXiv:2402.05129*.

Petter Törnberg. 2024b. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, page 08944393241286471.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-llm collaborative annotation through effective verification of llm labels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–21.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

A More Quantitative Analysis Examples

A.1 Defensible Divergence

Some other examples of defensible divergence in the dataset are

- **Text:***"The unfortunate Dowden suffers from memory loss."*
  - **LLM annotation:** Propaganda(Name Calling / Labeling)
  - **Justification**: The reasoning for selecting 'name calling or labeling' is that the sentence labels Dowden as 'unfortunate' and suffering from 'memory loss'. The use of 'unfortunate' implies a negative connotation, thus labeling Dowden in a potentially undesirable way to influence perceptions.
- **Text:***"Sorry, man—we're not giving up cake no matter the nuance."*
  - **LLM annotation:** Propaganda(Thought-Terminating Cliché)
  - **Justification**: The phrase 'no matter the nuance' suggests a refusal to consider any complexities or nuances related to the idea of giving up cake. This fits with the definition of a thought-terminating cliché, as it dismisses further discussion or critical thought on the topic by simplifying the decision to an absolute dismissal.

A.2 Reasonable but Overreaching

More examples of reasonable but overreaching ones are

- **Text:***"While Democratize SSMU later apologized for the language and deleted it from its site, the damage had already been done, Boudreau concluded."*
  - **LLM annotation:** Propaganda(Minimization or Exaggeration)
  - **Justification**: The phrase 'the damage had already been done' uses minimization or exaggeration to present the consequences of the original language (which was deleted and apologized for) as significant and irreversible, possibly overstating the impact without providing evidence for the extent of the damage.
- **Text:***"For those heavy into it, that's their thing, but it ain't my thing."*
  - **LLM annotation:** Propaganda(Thought-Terminating Cliché)
  - **Justification**: The phrase is a generic statement that simplifies the situation and shuts down further discussion by accepting others' involvement without engaging in deeper exploration or critique.

A.3 Unconvincing or Unsupported Reasoning

- **Text:***"All they have to do is delay the confirmation until after the election."*
  - **LLM annotation:** Propaganda(Oversimplification)
  - **Justification**: The sentence implies that a single, simple action—delaying the confirmation—is sufficient to achieve the desired outcome, thereby reducing a complex political process with multiple variables to one straightforward cause-and-effect step
- **Text:***"Each Carmel has its own spirit, its own tempo, its own feel."*
  - **LLM annotation:** Propaganda(Loaded Language, Repetition)
  - **Justification**: Words such as 'spirit', 'tempo', 'feel', and the emotive notion of a personal 'call to a specific Carmel' carry spiritually charged connotations meant to inspire or move the reader. The phrase 'its own' is repeated three times to reinforce uniqueness.

B Prompt

Annotation prompt: "You are an annotator for a neutral corporation that provides unbiased annotation services. You remain impartial to any political or ideological stance, such as left, right, or center. Your task is to detect text-based propaganda objectively, strictly adhering to the provided definition without introducing any personal or ideological bias."

Consolidation instruction: "You are a consolidator for a neutral corporation that provides unbiased annotation after reading annotations and reasoning by three annotators and definition of propaganda techniques. You remain impartial to any political or ideological stance, such as left, right, or center. As a subject matter expert, your job is to detect text-based propaganda, objectively, without introducing any personal or ideological bias. As an expert you are free to add or remove relevant or irrelevant techniques respectively based on the definitions. The annotators have worked individually based on the same definitions. Try to understand the view of annotators but be critical in your judgment before making the decision."

Definitions of rhetorical techniques: "Below is a list of rhetorical techniques with their definitions:

- loaded language: Using words/phrases with strong emotional implications (positive or negative) to influence an audience.
- name calling or labeling: Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or otherwise loves or praises.
- repetition: Repeating the same message over and over again, so that the audience will eventually accept it.
- exaggeration or minimization: Either representing something in an excessive manner: making things larger, better, worse.
- doubt: Questioning the credibility of someone or something. - appeal to fear/prejudice: Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative, possibly based on preconceived judgments.
- flag-waving: Playing on strong national feeling (or with respect to a group, e.g., race, gender, political preference) to justify or promote an action or idea.
- causal oversimplification: Assuming one cause when there are multiple causes behind an issue. We include scapegoating as well which is defined as the transfer of the blame to one person or group of people without investigating the complexities of an issue.
- slogans: A brief and striking phrase that may include labeling and stereotyping.
- appeal to authority: Stating that a claim is true simply because a valid authority/expert on the issue supports it, without any other supporting evidence. Include the special case where the reference is not an authority/expert,although it is referred to as testimonial in the literature.
- black-and-white fallacy: Presenting two alternative options as the only possibilities, when in fact more possibilities exist, eliminating any other possible choice. and as an extreme telling the audience exactly what actions to take, which is also called as dictatorship.
- thought-terminating cliche: Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are

typically short, generic sentences that offer seemingly simple answers to complex questions or that distract attention away from other lines of thought.

- whataboutism: Discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.
- reductio ad hitlerum: Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience. It can refer to any person or concept with a negative connotation.
- red herring: Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made.
- bandwagon: Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action".
- obfuscation/intentional vagueness/confusion: Using deliberately unclear words, so that the audience may have its own interpretation."

C Performance Analysis - Precision/Recall

The following table shows the class-wise precision and recall of both propaganda and non propaganda class as Precision(propaganda), recall (propaganda), precision (non-propaganda), and recall(non-propaganda). These tables shows the per-class performance of P/R/F results shown in Table 1– 4.

Table 7

Model	PTC P/R (prop, non-prop)	JMBX P/R (prop, non-prop)
GPT 4.1	0.68/0.91/0.87/0.58	0.75/0.79/0.78/0.74
GPT 4o	0.68/0.94/0.90/0.55	**0.78/0.73/0.75/0.80**
GPT o3	0.64/0.96/0.92/0.45	0.66/0.86/0.80/0.56

Table 7: GPT model performance across precision and recall for propaganda and non-propaganda classes on PTC and JMBX datasets. Values are *Precision/Recall* shown as *(prop, non-prop)*.

The following table 8shows per-class precision and recall for GPT models as **consolidators** in a multi-annotator setting.

Model	PTC P/R (prop, non-prop)	JMBX P/R (prop, non-prop)
GPT 4.1	0.64/0.96/0.92/0.47	0.67/0.86/0.81/0.58
GPT 4o	0.63/0.97/0.93/0.43	0.66/0.88/0.82/0.55
GPT o3	0.64/0.96/0.92/0.46	0.68/0.90/0.85/0.58

Table 8: GPT model performance across precision and recall for propaganda and non-propaganda classes on PTC and JMBX datasets. Values are *Precision/Recall* shown as *(prop, non-prop)*.

The following table for open source model per class precision and recall.

Model	PTC P/R (prop, non-prop)	JMBX P/R (prop, non-prop)
Qwen3-32B	**0.70/0.90/0.86/0.61**	**0.72/0.67/0.69/0.73**
QwQ-32B	0.63/0.97/0.94/0.44	0.60/0.90/0.80/0.40
DeepSeek R1	0.68/0.95/0.92/0.56	0.65/0.81/0.75/0.56

Table 9: Open-source LLMs performance across precision and recall for propaganda and non-propaganda classes on PTC and JMBX datasets. Values are *Precision/Recall* shown as *(prop, non-prop)*.

The following table 8shows per-class precision and recall for GPT models as **consolidators** in a multi-annotator setting.

Model	PTC P/R (prop, non-prop)	JMBX P/R (prop, non-prop)
Qwen3-32B(c)	0.66/0.95/0.91/0.52	0.66/0.81/0.75/0.58
QwQ-32B(c)	0.65/0.96/0.92/0.49	0.63/0.89/0.81/0.48
DeepSeek R1(c)	0.66/0.96/0.93/0.50	0.62/0.83/0.74/0.49

Table 10: Open-source LLMs performance across precision and recall for propaganda and non-propaganda classes on PTC and JMBX datasets. Values are *Precision/Recall* shown as *(prop, non-prop)*.