

Reasoning for Propaganda Detection via Label-Guided CoT Optimisation

Zeba Afroz, Harsh Vardhan, Md. Shad Akhtar
Indraprastha Institute of Information Technology Delhi
{zebaa, harsh25001, shad.akhtar}@iiitd.ac.in

Abstract

Propaganda detection systems usually operate as black boxes, lacking explanations. To address this, we hypothesise that persuasive techniques contain definitional characteristics with respect to their social-political setting. We introduce ReCoT, a reasoning-conditioned detection framework that captures definitional cues of persuasive techniques via population-level refinement and performance-guided re-sampling, without requiring additional explanation annotations. On INDI-PROPspan, ReCoT achieves a macro $F1_{LCS}$ of 0.364, surpassing baselines and reaching 88% of the SFT upper bound on Avg $F1_{LCS}$. Similarly, on PTC, ReCoT reaches 85% of the SFT upper bound on Avg $F1_{LCS}$, closing much of this gap while remaining interpretable. Importantly, the analysis of the reasoning pool’s convergence and update pattern for ReCoT shows effective refinement of reasoning CoTs.

1 Introduction

Propaganda aims to influence public opinion and create ideological biases through communicative strategies that can take textual, visual, or other forms. Unlike misinformation, propaganda often works with subtle cues to persuade. Persuasive techniques are one of the subtle strategies. Repeated exposure to these techniques can alter ideologies and normalise biased viewpoints. Hence, identifying these subtle persuasion techniques becomes crucial. Early research on propaganda detection mainly focused on basic classification tasks. The goal was to determine if a news article or paragraph contains propaganda (Rashkin et al., 2017; Barrón-Cedeno et al., 2019). Building on these, later research conducted fine-grained propaganda analysis and introduced technique-wise span-level annotations (Da San Martino et al., 2019). Further, the research advanced to multimodal and multilingual setups, as seen in SemEval workshops

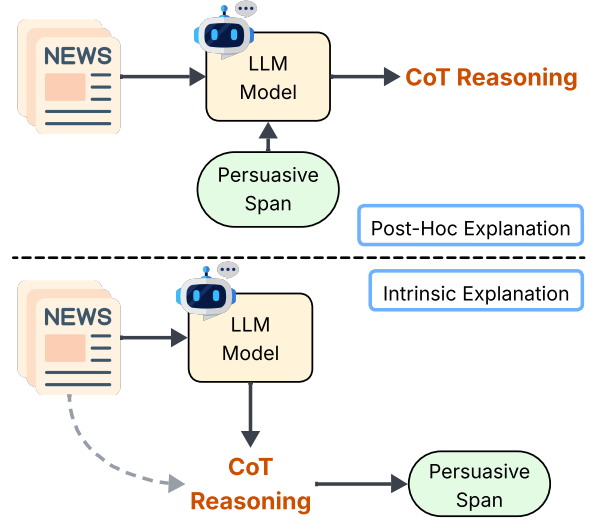


Figure 1: Conceptual comparison of Post-Hoc vs. Intrinsic Explanation frameworks. In the Post-Hoc setup (top), the Persuasive Span influences the LLM’s CoT Reasoning. In the Intrinsic setup (bottom), the LLM generates CoT Reasoning directly from the input news, which then identifies the relevant Persuasive Span.

(Da San Martino et al., 2020; Dimitrov et al., 2021, 2024). But we contend that merely identifying the specific technique is not enough, as earlier studies (Durmus and Cardie, 2019) have shown that readers often view information through their existing beliefs and ideological backgrounds. When propaganda aligns with a reader’s viewpoint, they might see it as neutral or factual rather than manipulative. Thus, merely pointing out persuasion is not enough; an understandable explanation must accompany the detection to provide a more critical perspective for the user, enabling them to engage in critical thinking and form sound, unbiased views.

This explainability gap is addressed by Yu et al. (2021) by proposing a model that detects propaganda based on syntactic and semantic features. And, Zavolokina et al. (2024) demonstrated in a user study that GPT-4-generated explanations can

encourage critical thinking and increase user trust in automated systems. Further, building on these, Hasanain et al. (2025) introduced a dataset for propaganda study in Arabic and English with explanation annotations. However, the annotations were done using post hoc reasoning, which doesn't causally guide the detection. We argue that for an explanation to be trustworthy, the reasoning must condition the decision process, ensuring an alignment between detection and its explanation. To address this, we hypothesise that persuasive techniques possess definitional cues with respect to their socio-political context. If a model can capture these cues, it can reason using them to identify persuasion, making explanation a causal part of the detection, as shown in Figure 1. This reasoning explains the detection.

Motivated by this intuition, we propose ReCoT, a reasoning-driven framework that first generates Chain-of-Thought (CoT) explanations and subsequently derives predictions conditioned on these explanations. To improve the quality of these reasoning chains (CoTs), we employ an iterative refinement strategy based on population-level cross-generalisation (Xu et al., 2023), in which CoTs from multiple instances of the same persuasion technique are exchanged and reused as few-shot examples during refinement. This process encourages the CoT to capture recurring cues associated with each technique. Consequently, CoTs progressively capture all necessary cues of persuasion techniques, enabling the framework to not only identify propaganda spans but also to provide context-aware and technique-aware explanations grounded in the reasoning process itself. To ensure that captured recurring definitional cues are related to the technique and are useful, the refinement process is guided by a sampling strategy that prioritises CoTs that consistently yield measurable performance improvements across iterations. This is shown, via update-density analysis, which shows convergence of the global pool.

Furthermore, we extend the existing INDI-PROP dataset (Afroz et al., 2026) with span-level annotations for persuasion techniques. As most existing data on span-level propaganda comes from Western and Arabic contexts, this overlooks the socio-political contexts of countries like India. To this end, we manually annotate 3922 spans across 20 persuasion techniques in 1266 news articles.

Contribution: Our contributions are as follows:

- We propose **ReCoT**¹, a reasoning-driven approach that employs population-based CoT sharing with iterative resampling for fine-grained propaganda detection.
- We introduce a bias-guided sampling strategy to utilize and refine reasoning over n iterations.
- We extend the existing INDI-PROP dataset with 3,922 span-level annotations on 20 persuasion techniques and named it INDI-PROPspan.

2 Related work

Early research on propaganda detection was at the document level. Rashkin et al. (2017) introduced one of the earliest datasets by compiling news articles from eight publishers and categorising them into propaganda, reliable, hoax, and satire. Building on this, Barrón-Cedeno et al. (2019) created a binary version by collapsing all non-propaganda categories into a single class. These early experiments were based on distant supervision, assuming that all articles from a given outlet share the same label. This is a highly noisy approach (Horne et al., 2018). Further Habernal et al. (2018b, 2017, 2018a) examined reasoning patterns and logical fallacies associated with persuasive discourse and introduced datasets targeting ad hominem and other fallacious reasoning. Although these efforts provided valuable insights into persuasive language, they did not fall within the domain of propaganda. A major advancement was made by Da San Martino et al. (2019) specifically for propaganda detection in news, which proposed a fine-grained annotation covering 18 persuasive techniques at the span level, reframing propaganda detection as a span-extraction task. This formulation served as the basis for the SemEval-2020 shared task (Da San Martino et al., 2020). Then, subsequent SemEval editions expanded the task to multimodal settings using memes (Dimitrov et al., 2021, 2024) to multilingual scenarios (Piskorski et al., 2023). Despite these advances, most existing systems focus on identifying where propaganda occurs rather than explaining why a span is propagandistic. These detection systems were not interpretable. To address this gap, Yu et al. (2021) proposed interpretable approaches that integrate syntactic and semantic features for propaganda detection. Recently, Zavolokina et al. (2024) demonstrated in a user study that GPT-4-generated explanations can encourage crit-

¹Code and the subset of the dataset are available at https://anonymous.4open.science/r/Prop_Span-6FF7

ical thinking and increase user trust in automated systems. Building on these efforts, Hasanain et al. (2025) introduced the first explanation-integrated dataset for propaganda detection in Arabic and English. Although this dataset represents an important step toward explainable propaganda detection, the explanations are generated post hoc. In contrast, our work focuses on generating intrinsic explanations alongside span-level predictions. As we argue that, prediction must be causal to reasoning. Specifically, we leverage LLMs and CoT prompting to generate reasoning without additional annotation overhead, and refine them iteratively to improve the quality and consistency of these CoTs. In addition, recent refinement methods such as Self-Refine (Madaan et al., 2023), Reflexion (Shinn et al., 2023), and Self-Consistency (Wei et al., 2022) demonstrate the benefits of iterative refinement but rely primarily on self-feedback, memory, or ensembling. Xu et al. (2023) extends this paradigm by introducing a population-based prompt refinement strategy, and Guo et al. (2023) is based on evolutionary algorithms. Building on this line of work, we incorporate cues of persuasion techniques to guide convergence during population-based refinement, retaining only refinements that demonstrate measurable improvements. This refinement process is implemented through resampled in-context learning.

3 Dataset

We build upon the dataset introduced by Afroz et al. (2026), which comprises 1,266 news articles covering two major socio-political events in India. We annotated each article in the INDI-PROP corpus with span-level labels for the existing persuasive techniques, following the definitions established by the original paper. The final corpus contains 3,922 annotated spans across 20 persuasive techniques. And we name it **INDI-PROPspan**.

3.1 Annotation

To ensure annotation quality, we recruit annotators from diverse age ranges, ideological stances, and gender identities. In total, 9 annotators participated, aged 20-40 years, comprising 4 women and 5 men. The annotation team included research assistants, postgraduate, graduate students, and industry researchers. Additionally, two moderators were appointed to oversee and ensure the consistency of the annotations. All annotators and mod-

ID	Technique	Train	Test	Total
T1	Causal Oversimplification	99	21	120
T2	Black-and-white Fallacy	28	4	32
T3	Straw Man	55	17	72
T4	Whataboutism	32	10	42
T5	Reductio ad Hitlerum	22	6	28
T6	Red Herring	31	13	44
T7	Loaded Language	273	69	342
T8	Name Calling and Labeling	235	67	302
T9	Appeal to emotion/fear/prejudice	411	104	515
T10	Slogans	164	41	205
T11	Flag-Waving	77	21	98
T12	Exaggeration and Minimization	56	12	68
T13	Thought-terminating Clichés	85	18	103
T14	Bandwagon	44	8	52
T15	Smears	671	180	851
T16	Obfuscation, Vagueness, Confusion	29	6	35
T17	Doubt	124	25	149
T18	Appeal to Authority	58	14	72
T19	Assertion	449	112	561
T20	Glittering Generalities (Virtue)	100	28	128
Total	All Techniques	3126	796	3922

Table 1: Frequency distribution of propaganda techniques in INDI-PROPspan.

erators received compensation in accordance with institutional guidelines.

Annotator Training: The training process was organised in iterative stages to ensure conceptual alignment and consistency among annotators.

1. **Orientation & Pilot Annotation Phase:** The training began with an orientation session. Annotators were thoroughly familiarised with the annotation guidelines, which detailed each persuasive technique through definitions and examples. Following the orientation, annotators independently annotated an initial pilot set of articles. This pilot dataset included ambiguous cases to reveal common sources of confusion and interpretation variation at an early stage.
2. **Annotation Refinement:** After each annotation round, meetings were conducted to compare annotations and discuss the rationale behind differing labels systematically. Inter-annotator agreement (IAA) metrics were computed after each cycle for quantitative assessment. Based on these outcomes, guidelines were refined, and discussions were iterated until satisfactory IAA scores were achieved across annotators.

Final IAA Score: For inter-annotator agreement (IAA), the pairwise overlap-based F1 score is used. The span overlap is calculated using the longest common substring method. The final inter-annotator agreement is calculated as the average across all annotator pairs. The IAA is 0.63, calculated for 20 articles containing 46 spans. Measuring span-level agreement is challenging as minor boundary differences reduce agreement. Moreover,

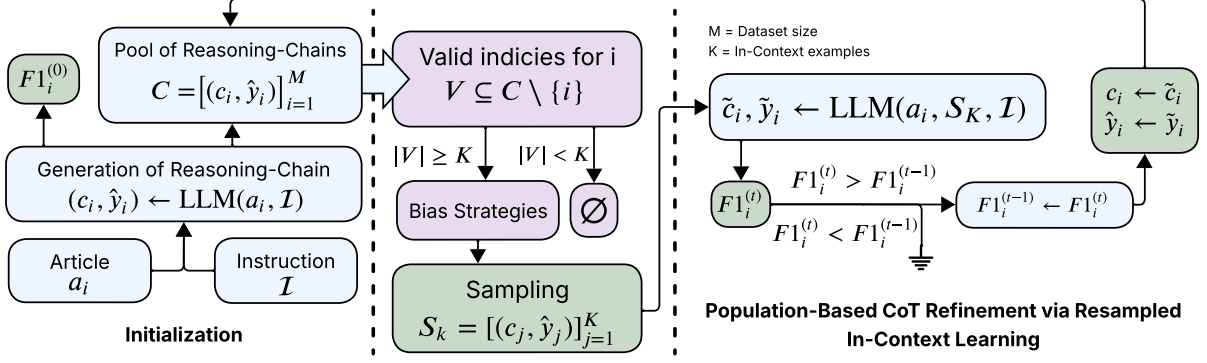


Figure 2: Overview of the proposed ReCoT framework. Block 1: illustrates the initialisation stage, where initial CoT-predictions pairs are generated, Block 2: presents the sampling process & Block 3: shows the iterative population-based refinement process via resampled in-context learning.

the evaluation subset deliberately included ambiguous and borderline cases to stress-test annotator alignment; it represents a harder-than-average set of cases rather than a random, representative sample.

3.2 Dataset Statistics

The distribution of persuasion techniques across INDI-PROPspan shows that Smears and Assertion are the most frequent, accounting for approximately 21.7% and 14.3% of all persuasive spans, respectively, reflecting the prevalence of accusatory and fact-like strategies in the corpus. In contrast, Reductio ad Hitlerum is the least frequent technique, accounting for less than 1%. Exact counts for each technique are shown in Table 1, and a detailed statistical discussion is in the Appendix B.

4 Methodology

In contrast to the post-hoc method, where the explanation is generated as a secondary process to the predictions, our framework directly conditions span extraction on CoT Reasoning. This allows the reasoning’s performance to directly influence detection. We hypothesise that each persuasive technique has definitional cues shaped by the context in which it is implied, and to enable the CoTs to capture these cues, the CoTs were iteratively refined through population-level interactions. During refinement, only reasoning chains (i.e., CoTs) that improve span-level performance according to an acceptance criterion are retained, progressively reinforcing technique-relevant cues, as illustrated in Figure 2. ReCoT proceeds in two stages: (1) initialization, and (2) iterative refinement. Each article a_i in the dataset $\mathcal{T} = \{a_1, a_2, \dots, a_M\}$ is annotated

with persuasive technique spans y_i .

Initialization: For each article a_i with ground-truth labels y_i , we obtain an initial chain-of-thought reasoning c_i and a corresponding prediction $\hat{y}_i$ using LLM in zero-shot, given instruction $\mathcal{I}$. Refer to Figure 2. The pair $(c_i, \hat{y}_i)$ forms the initial reasoning chain for article a_i , which serves as the starting point for subsequent refinement steps. From the initialization step, we obtain the list C (Global pool), which contains all initial CoTs (Reasoning chains) for the dataset of M articles.

Sampling (I-Bias): We define a biased sampling strategy that favours CoTs, which have been improved in previous iterations. The primary goal is to maximise the reuse of refined reasoning. We keep δ (the bias) small and renormalise it to maintain an exploratory nature during sampling. We maintain a global probability array $\mathbf{p} \in \mathbb{R}^M$, initialised uniformly at the beginning for all dataset indices. At each iteration, we aim to refine the CoT at a current index $i \in \{1, \dots, M\}$. After each refinement step, if the CoT at index i yields a higher $F1_{LCS}$ score than its previous version, CoT is updated, and the probability of index i is incremented by a fixed bias $\delta = 0.02$, subject to a maximum cap of $p_{\max} = 0.1$, illustrated as $\hat{p}_i^{(t)} = \min(p_i^{(t-1)} + \delta, p_{\max})$. This will increase the likelihood of sampling the improved CoTs. For sampling CoTs, we maintain a subset $V \subseteq C \setminus \{i\}$, which denotes the set of valid indices corresponding to non-empty CoTs. i , denoting the current index, is excluded from the candidate pool of valid indices to avoid self-conditioning.

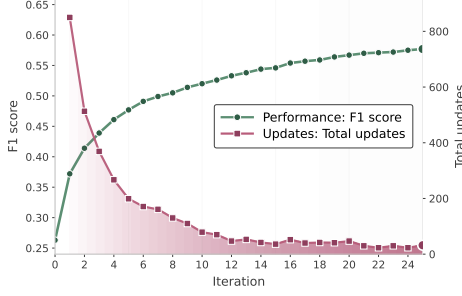


Figure 3: ReCoT’s refinement and update density convergence curve across 25 iterations for i -bias & $k=2$ configuration on the INDI-PROPspan dataset

$$f(i, K) = \begin{cases} \text{Sample}(V, K, \mathbf{p}_V^{(t)}), & \text{if } |V| \geq K \\ \emptyset, & \text{if } |V| < K \end{cases} \quad (1)$$

After increasing the bias, the global probability array is then renormalised to maintain a valid distribution. The sampling function $f(i, K)$ then draws K support indices (K in-context examples) from V as shown in equation (1), where $\mathbf{p}_V^{(t)}$ is the renormalized distribution restricted to V . If instances in the valid set are less than the value of K , which is the required number of in-context examples, then in this scenario, a null set is passed. This bias strategy ensures that CoTs that have shown consistent improvement are more frequently sampled as in-context examples, allowing the model to leverage higher-quality reasoning chains during refinement.

Iterative Refinement: After initialization, we iteratively refine each article’s reasoning chain (CoT) using in-context learning (Brown et al., 2020), thereby improving its predictions. In each refinement step for a given instance, we sample CoTs from C to construct K -shot in-context examples using the I-bias strategy. These examples are fed into the LLM along with the target article a_i and instruction $\mathcal{I}$ to produce new refined reasoning chains and predictions conditioned on the examples provided. Illustrated as $(\tilde{c}_i, \tilde{y}_i) \leftarrow \text{LLM}(a_i, S_K, \mathcal{I})$, here S_K is the sampled in-context examples of size K . If the new prediction achieves a higher $F1_{LCS}$ score compared to the previous version for a given instance i , we accept the updated pair $(\tilde{c}_i, \tilde{y}_i)$ and replace the old one, as shown in Figure 2. This process is repeated across the entire dataset and for multiple iterations.

Inference: After refinement, R examples are sampled from the evolved global pool and inserted

Config	Backbone	IndiProp		PTC	
		Avg	W.Avg	Avg	W.Avg
SFT (Upper Bound)	Mistral	0.412	0.538	0.235	0.266
Hitachi	—	0.180	0.232	0.093	0.118
Aschern		0.205	0.231	0.174	0.203
0-shot	Mistral	0.263	0.273	0.187	0.184
1-shot		0.318	0.317	0.200	0.192
2-shot		0.154	0.157	0.195	0.194
CoT		0.285	0.300	0.186	0.202
RePrompt	Mistral	0.326	0.314	0.160	0.186
EvoPrompt		0.213	0.227	0.171	0.195
GEPA		0.296	0.324	0.233	0.278
0-shot	GPT	0.294	0.279	0.169	0.163
1-shot		0.254	0.263	0.218	0.215
2-shot		0.275	0.279	0.212	0.203
CoT		0.302	0.289	0.210	0.208
RePrompt	GPT	0.297	0.308	0.193	0.209
EvoPrompt		0.272	0.300	0.232	0.208
ReCoT R3	Mistral	0.364	0.347	0.155	0.172
ReCoT R3	Qwen	0.345	0.362	0.184	0.198
ReCoT R4		0.326	0.356	0.200	0.211
ReCoT R3	GPT	0.320	0.216	0.165	0.130

Table 2: $F1_{LCS}$ Performance comparison on INDI-PROPspan and PTC for i -bias & $K=2$ setup.

into the prompt as few-shot examples, to produce predicted spans and reasoning for the test article, evaluated using the metric in §5.

5 Experiments, Results, and Analyses

We use Mistral-Small-3.2-24B-Instruct for both initialization and refinement, for $n = 25$ refinement iterations with $K \in \{1, 2, 3\}$ in-context example. The temperature is set to 1.0 during both refinement and inference. The prompts used are provided in Appendix C. We evaluate three inference backbones: Mistral-Small-3.2-24B-Instruct, Qwen3-32B, and GPT-4o-mini. To study the effect of the number of in-context examples during inference, we experiment with five setups where $R \in \{1, 2, 3, 4, 5\}$. The refinement and inference are conducted on the proposed INDI-PROPspan and PTC² (Da San Martino et al., 2020), and the reasoning pool is refined independently on both datasets. The average inference time per instance during inference is detailed in Appendix F. We also fine-tuned Mistral-Small-3.2-24B-Instruct (SFT) as an upper-bound reference.

Baselines: We compared against **SemEval Models:** Aschern (Chernyavskiy et al., 2020) and Hitachi (Morio et al., 2020), **Prompting baselines:** Zero, One, Two-shot, and CoT (Wei et al., 2022) settings. **Reasoning baselines:** the RePrompting

²A widely used benchmark for propaganda detection introduced in SemEval-2020 Task 11 for research and academic purposes.

Setup	R	Mistral				Qwen			
		IndiProp		PTC		IndiProp		PTC	
		Avg	W.Avg	Avg	W.Avg	Avg	W.Avg	Avg	W.Avg
$K = 1$	R=1	0.323	0.336	0.171	0.185	0.323	0.339	0.184	0.201
	R=2	0.340	0.337	0.171	0.185	0.313	0.344	0.190	0.205
	R=3	0.354	0.360	0.182	0.196	0.321	0.358	0.187	0.199
	R=4	0.334	0.345	0.193	0.205	0.341	0.352	0.177	0.192
	R=5	0.357	0.360	0.186	0.196	0.323	0.342	0.170	0.185
$K = 2$	R=1	0.320	0.311	0.167	0.180	0.296	0.326	0.172	0.191
	R=2	0.343	0.300	0.166	0.182	0.316	0.348	0.181	0.196
	R=3	0.364	0.347	0.155	0.172	0.345	0.362	0.184	0.198
	R=4	0.344	0.317	0.172	0.170	0.326	0.356	0.200	0.211
	R=5	0.361	0.306	0.172	0.170	0.359	0.355	0.170	0.187
$K = 3$	R=1	0.347	0.333	0.156	0.153	0.330	0.343	0.194	0.204
	R=2	0.319	0.340	0.157	0.157	0.310	0.343	0.183	0.199
	R=3	0.334	0.338	0.168	0.165	0.340	0.355	0.178	0.192
	R=4	0.324	0.331	0.153	0.152	0.326	0.343	0.166	0.182
	R=5	0.328	0.347	0.156	0.156	0.343	0.353	0.180	0.183

Table 3: ReCoT Performance across different values of K using i_{bias} strategy

algorithm (Xu et al., 2023), a population-based refinement baseline that employs Gibbs sampling, and EvoPrompt (Guo et al., 2023), based on evolutionary algorithms. and for a more recent baseline, we also included GEPA (Agrawal et al., 2026)

Evaluation Metrics: We adopt evaluation metrics based on the longest common subsequence, given in equations 2 and 3, where S_{pred} denotes predicted spans and S_{gt} denotes ground truth spans. To compute $F1_{LCS}$ we take harmonic mean of P_{LCS} and R_{LCS} . $F1_{LCS}$ evaluates both refinement and inference.

$$P_{LCS} = \frac{1}{|S_{pred}|} \sum_{p \in S_{pred}} \max_{g \in S_{gt}} \left(\frac{LCS(p, g)}{|p|} \right) \quad (2)$$

$$R_{LCS} = \frac{1}{|S_{gt}|} \sum_{g \in S_{gt}} \max_{p \in S_{pred}} \left(\frac{LCS(p, g)}{|g|} \right) \quad (3)$$

Early iterations might result in a partially correct answer, an exact match criterion will reject the CoTs of early iterations, eliminating the learning cue, whereas $F1_{LCS}$ rewards partial token overlap, allowing the pool to progress gradually from a few correct tokens toward full span coverage across iterations.

Global Pool Convergence: Figure 3 shows the refinement of the global reasoning pool C . The figure plots the relationship between update density and $F1_{LCS}$ score across 25 iterations on the Mistral backbone for the INDI-PROPspan dataset (PTC convergence curve is given in Appendix D). The refinement curve exhibits a two-phase trajectory: an initial phase of rapid performance improvement followed by convergence in later iterations. In the

early stages of refinement, the pool contains noisy reasoning chains. Consequently, this results in a large number of updates. Through population-level cross-example generalisation and I-bias strategy, ReCoT progressively reinforces reasoning chains that capture useful cues. As refinement progresses, the performance gains gradually saturate, indicating that the reasoning pool has stabilised and that further updates provide only limited improvement. Consequently, the number of accepted updates also decreases in later iterations. Representative examples comparing zero-shot and refined CoTs are presented in Appendix G for the comparison of refined CoT with initialisation. The zero-shot CoT often relies on generic reasoning and extracts broader or less relevant spans. After iterative refinement, the CoTs become more focused on the defining characteristics of the target technique, resulting in more precise evidence selection.

Comparative Results: Table 2 presents the quantitative analysis. Comparing with SFT (Upper bound reference), ReCoT $R = 3$ (Mistral) on INDI-PROPspan achieves 88% of the SFT upper bound on Avg $F1_{LCS}$ and 64% on W.Avg $F1_{LCS}$. The larger gap on the weighted metric is consistent with SFT’s advantage on high-frequency techniques, which benefit from direct gradient-based adaptation. Similarly, on PTC $R = 4$ (Qwen), ReCoT reaches 85% of the SFT upper bound on Avg $F1_{LCS}$ and 79% on W.Avg $F1_{LCS}$, closing much of this gap while remaining interpretable.

INDI-PROPspan with Mistral and Qwen at $R=3$ achieves the best macro $F1_{LCS}$ of 0.364 and weighted $F1_{LCS}$ of 0.362, respectively. These results surpass all prompting and reasoning baselines. Mistral configuration improves macro $F1_{LCS}$ by +0.101, +0.046 and +0.038 over zero, one-shot and RePrompting, respectively. On PTC, ReCoT delivers comparable rather than superior results relative to the baselines; the best ReCoT’s configuration (Qwen, $K = 2$, $R = 4$) achieves 0.200 average and 0.211 weighted $F1_{LCS}$. We hypothesise that, in the PTC dataset, ReCoT’s performance degradation is driven by domain/topic diversity in the in-context examples selected during inference. INDI-PROPspan is drawn from two closely related, temporally concentrated events (CAA and the farmers’ protest), so examples sampled from the reasoning pool at inference time share the same domain and framing, even though selection is random. PTC spans many unrelated events, so random sampling

at inference more often pulls in-context examples from a different topic/domain than the test instance. This shows context/domain dependence of ReCoT (A detailed explanation is provided in Appendix D). We illustrate this with representative cases below:

- PTC - [Article: Glamorising (US Military) + R=3 in-context examples: Neutral reporting (Iran regime) + criticise(American President)+ Report(Plague in Madagascar), showing mismatched domains]
- INDI-PROPspan - [Article: criticise(CAA protest) + R=3 in-context examples: criticise(CAA protest) + criticise(Farmers' protest) + criticise(Government over CAA bill), showing same-event/domain alignment]

Overall, $R=3$ & $R=4$ perform better than other configurations, as a smaller R provides insufficient in-context examples and a larger R introduces more noise. R_{LCS} and P_{LCS} for $K=2$ configuration is reported in Appendix F along with the quantitative results for $K=1$ and $K=3$ configurations. We also evaluated using the Sem-Eval metrics introduced in Da San Martino et al. (2019), the results for the Mistral backbone are given in Table 4. Full table is provided in the Appendix³ F, On INDI-PROPspan, ReCoT outperforms both SemEval models and standard prompting baselines on both backbones, peaking at avg. F1 0.277 ($R=3$) / w.F1 0.280 ($R=2$). On PTC, scores are lower overall, with ReCoT peaking at avg. F1 0.120 / w.F1 0.164 ($R=2$). These results mirror the trend observed with our LCS-based metric: PTC scores are lower across all methods, including SFT, reflecting the benchmark’s intrinsically difficult nature.

Config	Backbone	IndiProp		PTC	
		Avg	W.Avg	Avg	W.Avg
SFT (Upper Bound)	Mistral	0.210	0.214	0.127	0.195
Hitachi	—	0.161	0.207	0.092	0.143
Aschern		0.152	0.193	0.089	0.141
GEPA		0.247	0.256	0.158	0.185
0-shot	Mistral	0.201	0.213	0.111	0.139
1-shot		0.244	0.242	0.159	0.158
CoT		0.213	0.235	0.109	0.148
RePrompt	Mistral	0.263	0.259	0.147	0.168
ReCoT R1		0.222	0.252	0.098	0.144
ReCoT R2		0.272	0.280	0.120	0.164
ReCoT R3		0.277	0.276	0.093	0.151
ReCoT R4		0.258	0.258	0.106	0.155
ReCoT R5		0.268	0.277	0.103	0.135

Table 4: Performance comparison on the SemEval metrics - Avg F1 and Weighted Average (W.Avg) F1 are reported for the IndiProp and PTC tasks.

³Due to the brevity, only the Mistral backbone is reported here in Table 4.

Category of error	Remark / Error Analysis
GT: [ORG] ⁴ was misleading people by saying that [POLICY] is against minorities Pred: [ORG] was misleading people by saying that [POLICY] is against minorities, and asked them to show proof for that claim [Over-prediction]	Model added extra explanatory clause ("asked them to show proof"), extending beyond the gold boundary, another over-prediction error.
GT: The [ORG] is targeting a particular community by carrying out the [POLICY] in North and South 24 Pargana districts [Under-prediction] Pred: targeting a particular community	Model extracted only the core phrase, ignoring key qualifiers ("[ORG]", "[POLICY]"), leading to a truncated reasoning span.
GT: No Propaganda span present. Pred: what they did was 'unlawful' and 'not legally correct' [False Positive]	Model predicted a span even though no persuasive content was annotated. False trigger due to lexical cues like "unlawful" and "not legally correct".
GT: They won't be able to give citizenship to anyone in two days. This is just a lollipop and a show-off. [False Negative] Pred: No Propaganda span present.	Model failed to detect explicit ridicule and exaggeration, missing emotionally charged language recall failure on loaded or sarcastic tone.

Table 5: Representative examples of common error categories in span-level propaganda detection, including over-prediction, under-prediction, false positives, and false negatives.

Iter	Human Evaluation						Expert Evaluation					
	CR			RT			OC			TP		
	poor	fair	good	poor	fair	good	poor	fair	good	poor	fair	good
0	41	45	15	24	55	21	26	49	25	60	40	0
25	3	30	67	1	14	85	4	23	74	0	17	83

Table 6: Human and expert evaluation for assessing reasoning quality (in % values)

Ablation: We also report the effect of in-context examples during refinement, Table 3 analyses the effect of varying in-context examples $K \in \{1, 2, 3\}$ on the refinement process. $K=2$ consistently provides the strongest overall performance across both datasets, particularly for configurations with $R=3$ and $R=4$. Whereas $K=1$ provides a limited example for cross-example generalisation, restricting the model’s ability to accurately capture cues across persuasion techniques. In contrast, larger configurations, such as $K=3$, introduce noise, which negatively affects the refinement process. These results suggest that moderate population-level interaction during refinement is sufficient for capturing and technique-aware reasoning patterns. Whereas 7 reports cross-dataset transfer performance for the Mistral ($K=2$) configuration. On PTC, cross-dataset transfer performs comparably to, and in fact exceeds, same-dataset inference: average F1 ranges from 0.185 to 0.224 across R , compared to 0.155–0.173 when ReCoT is refined and evaluated on PTC alone (Table 2, $R=3$). On INDI-PROPspan, cross-dataset transfer shows a modest drop across R : average F1 ranges from 0.286 to 0.327, compared to 0.320–0.351 when ReCoT is both refined and evaluated on INDI-PROPspan itself (Mistral backbone, $R=3$).

R	PTC		IndiProp	
	Avg F1	W. Avg F1	Avg F1	W. Avg F1
1	0.188	0.196	0.286	0.270
2	0.209	0.207	0.317	0.295
3	0.185	0.189	0.327	0.299
4	0.194	0.190	0.315	0.286
5	0.224	0.205	0.306	0.279

Table 7: Cross-dataset inference performance for the Mistral (K2) configuration. PTC was evaluated using CoT refined on IndiProp, while IndiProp was evaluated using CoT refined on PTC.

These results suggest that CoT pools refined on one dataset transfer effectively to the other without requiring re-refinement, and in the case of PTC, even yield improved performance likely because the INDI-PROPspan-refined pool encodes more diverse, less domain-anchored reasoning cues that generalise better to PTC’s heterogeneous topics. The same-dataset comparison figures are drawn from the best-performing $R = 3$ configuration (Mistral) in Table 2. In addition to the proposed selection strategy, we evaluate different bias strategies to examine the effects of exploratory and exploitative biases during sampling and in the refinement process. A detailed discussion of bias strategies is in Appendix A and F.

Error Analysis: Error analysis reveals four failure modes: over-prediction, under-prediction, false positives, and false negatives, as shown in Table 5. Over-prediction: model captures excessive contextual text beyond the gold span, reducing precision. Under-prediction: arises from missing text, which impacts recall. False positives triggered by strong lexical cues (e.g., legal terms) in otherwise factual text. False negatives primarily involve subtle, implicit, or sarcastic propaganda, indicating difficulty in capturing nuanced pragmatic cues. These errors highlight the challenges of accurately detecting span boundaries.

5.1 Human analysis

To qualitatively assess the refined CoT explanations, we evaluate CoTs from iterations 0 and 25, 20 CoT instances were evaluated at iteration 0 and iteration 25, rated by 11 human evaluators and 9 expert evaluators. Human evaluators focused on interpretability-related factors. And experts trained in propaganda detection assessed CoTs for technical grounding, emphasising technique purity and logical structure. Each criterion defined for analy-

sis is scored as poor, fair, and good, as described in Table 6. To reduce evaluation bias, CoTs were anonymised, and Evaluators were not informed of the CoT’s origin.

Human Evaluation Criteria: Assess the interpretability and readability of refined CoTs, particularly how well the reasoning justifies the extracted information from the source text. **Clarity of Reasoning(CR):** Measures whether the CoT is clear and whether each step is easy to understand. **Relevance to the Text(RT):** Assesses whether the reasoning refers only to the article content and avoids hallucinations. It also judges whether the spans extracted by the CoT are actually present and semantically meaningful. **Overall Coherence(OC):** Rates whether the reasoning flows logically and is consistent with the final extracted spans.

Expert Evaluation Criteria: The expert evaluators are the same annotators who contributed to the dataset and developed expertise in identifying propaganda spans with logical justification. They evaluate the logical structure and technique purity of the generated reasoning. **Technique Purity (TP):** Determines whether the CoT adheres to the intended propaganda technique (e.g., Name-Calling/Labelling) without mixing unrelated techniques such as Appeal to Fear or Loaded Language. It also evaluates whether the CoT correctly explains and applies the technique’s definition. **Logical Structure (LS):** Assesses whether the reasoning is stepwise, methodical, and generalizable.

6 Conclusion

We present ReCoT, an interpretable propaganda detection framework that provides explanations for identifying propagandistic spans, unlike standard black-box approaches. ReCoT leverages population-level CoTs refinement, using a sampling strategy that prioritises CoTs that yield measurable performance gains and progressively captures reasoning cues. Along with interpretability, ReCoT reaches up to 88% and 85% of a supervised fine-tuning (SFT) upper bound on Avg $F1_{LCS}$ on INDI-PROPspan and PTC respectively. With competitive performance across all baselines on the INDI-PROPspan dataset, while PTC shows modest performance; lower scores across all methods, including SFT, reflect the benchmark’s intrinsically harder and more topic-diverse nature of PTC.

Limitations

While ReCoT shows promising results, it has several limitations. The homogeneous domain/event yields better results compared to heterogeneous domains. And the iterative refinement framework is computationally intensive, as it involves multiple refinement steps. Future work should explore lightweight refinement strategies, prompt compression, or smaller specialised models to improve efficiency. Errors or inaccuracies in text generated by large language models (LLMs) may affect the quality of the resulting reasoning chains. Currently, the framework is limited to English-language news articles; further studies can incorporate multilingual or code-mixed content.

Ethical Considerations

This work aims to promote transparency in media analysis by helping researchers and readers to detect persuasive and potentially manipulative communication patterns in news articles. But it is necessary to deploy propaganda-detection technologies responsibly and to read them with caution. Automated propaganda detection systems are not to be used to determine the truthfulness, legitimacy, or political validity of opinions expressed in the media. The proposed framework detects persuasive patterns and produces reasoning-based explanations, but these outputs are intended to support human analysis rather than to substitute for human judgment. Users might be vulnerable to over-trusting automatically generated explanations by large language models. As hallucinations and factually inconsistent reasoning are known issues with LLMs, the system’s explanations may sometimes contain inaccuracies or incomplete interpretations. So, human oversight and expert verification remain essential for interpreting the outputs. Propaganda detection systems might unintentionally reinforce societal or political biases embedded in the underlying datasets or language models. Therefore, such systems should be carefully evaluated in different domains, perspectives and demographic contexts before they are deployed in sensitive real-world applications. Finally, this research is not intended to promote censorship or restrict political speech. Instead, the proposed framework is intended to be an assistive analytical tool that fosters critical thinking and media literacy by helping users explore rhetorical and persuasive strategies in news discourse.

References

- Zeba Afroz, Harsh Vardhan, Pawan Bhakuni, Aanchal Punia, Rajdeep Kumar, and Md. Shad Akhtar. 2026. [Fine-grained narrative classification in biased news articles](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8431–8445, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2026. [Gepa: Reflective prompt evolution can outperform reinforcement learning](#). Preprint, arXiv:2507.19457.
- 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.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. 33:1877–1901.
- Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. 2020. [Aschern at SemEval-2020 task 11: It takes three to tango: RoBERTa, CRF, and transfer learning](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1462–1468, Barcelona (online). International Committee for Computational Linguistics.
- Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. [SemEval-2020 task 11: Detection of propaganda techniques in news articles](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.
- Giovanni Da San Martino, Yu Seunghak, Alberto Barrón-Cedeno, Rostislav Petrov, Preslav Nakov, and 1 others. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646. Association for Computational Linguistics.
- Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. [SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, Mexico City, Mexico. Association for Computational Linguistics.

- Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. [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.
- Esin Durmus and Claire Cardie. 2019. Exploring the role of prior beliefs for argument persuasion. *arXiv preprint arXiv:1906.11301*.
- Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujia Yang. 2023. Evoprompt: Connecting llms with evolutionary algorithms yields powerful prompt optimizers. *arXiv e-prints*, pages arXiv-2309.
- Ivan Habernal, Raffael Hannemann, Christian Polak, Christopher Klammer, Patrick Pauli, and Iryna Gurevych. 2017. [Argotario: Computational argumentation meets serious games](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.
- Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018a. [Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. [Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.
- Maram Hasanain, Md Arif Hasan, Mohamed Bayan Kmainasi, Elisa Sartori, Ali Ezzat Shahroor, Giovanni Da San Martino, and Firoj Alam. 2025. [PropXplain: Can LLMs enable explainable propaganda detection?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23855–23863, Suzhou, China. Association for Computational Linguistics.
- Benjamin Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.
- Gaku Morio, Terufumi Morishita, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. [Hitachi at SemEval-2020 task 11: An empirical study of pre-trained transformer family for propaganda detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1739–1748, Barcelona (online). International Committee for Computational Linguistics.
- 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.
- 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, Copenhagen, Denmark. Association for Computational Linguistics.
- Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflexion: Language agents with verbal reinforcement learning](#). *Preprint*, arXiv:2303.11366.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Weijia Xu, Andrzej Banburski-Fahey, and Nebojsa Jojic. 2023. Reprompting: Automated chain-of-thought prompt inference through gibbs sampling. *arXiv preprint arXiv:2305.09993*.
- Seunghak Yu, Giovanni Da San Martino, Mitra Mohtarami, James Glass, and Preslav Nakov. 2021. [Interpretable propaganda detection in news articles](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1597–1605, Held Online. INCOMA Ltd.
- Liudmila Zavolokina, Kilian Sprenkamp, Zoya Katashinskaya, Daniel Gordon Jones, and Gerhard Schwabe. 2024. Think fast, think slow, think critical: designing an automated propaganda detection tool. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–24.

A Bias Strategies

A.1 Random (No Bias)

We define sampling from the list C as follows. At each iteration, we aim to refine the CoT at a current index $i \in \{1, \dots, M\}$, which is excluded from the candidate pool to avoid self-conditioning.

Let $V \subseteq C \setminus \{i\}$ denote the set of valid indices corresponding to non-empty CoTs, excluding the index i . The sampling function $f(i, K)$ samples a subset with replacement, and each reasoning chain contains equal probability of selection. $S \subseteq V$ of size K , such that $|S| = K$, Where K is the number of in-context support examples. The formal definition is given in equation 4.

$$f(i, K) = \begin{cases} \text{Sample}(V, K), & \text{if } |V| \geq K \\ \emptyset, & \text{if } |V| < K \end{cases} \quad (4)$$

The sampling strategy ensures that each refinement step is based on a randomly selected set of CoTs and their corresponding prediction.

A.2 Sampling with Bias on Updating Index

We define a biased sampling strategy that favors CoTs which have previously contributed to the improvement of other reasoning chains. At each iteration, we aim to refine the CoT at a current index $i \in \{1, \dots, M\}$, which is excluded from the candidate pool to avoid self-conditioning.

Let $V \subseteq C \setminus \{i\}$ denote the set of valid indices corresponding to non-empty CoTs, excluding index i . We maintain a global probability array $\mathbf{p} \in \mathbb{R}^M$, initialized uniformly as:

$$p_j^{(0)} = \frac{1}{M}, \quad \forall j \in \{1, \dots, M\} \quad (5)$$

After each refinement step, if the CoT at index i yields a higher F1 score than its previous version, the probability of the support index $j \in S$ (S as defined in section A.1) that caused the update is incremented by a fixed bias $\delta = 0.02$, subject to a maximum cap of $p_{\max} = 0.1$:

$$\hat{p}_j^{(t)} = \min(p_j^{(t-1)} + \delta, p_{\max}) \quad (6)$$

The global probability array is then immediately renormalized to maintain a valid distribution:

$$p_k^{(t)} = \frac{\hat{p}_k^{(t)}}{\sum_{l=1}^M \hat{p}_l^{(t)}}, \quad \forall k \in \{1, \dots, M\} \quad (7)$$

where $\hat{p}_k^{(t)} = p_k^{(t-1)}$ for all $k \neq j$, i.e., only the support index j is modified before renormalization.

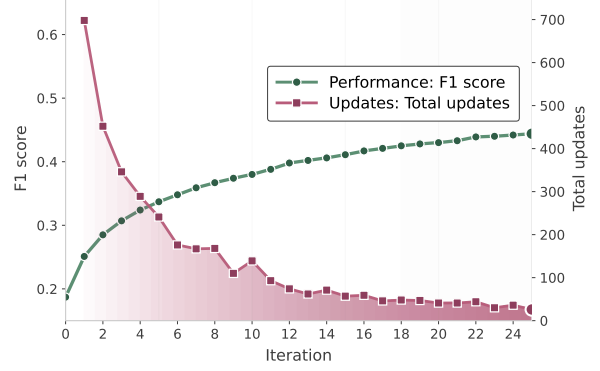


Figure 4: ReCoT’s refinement and update density convergence curve across 25 iterations for i -bias & $k=2$ configuration on the PTC dataset

The sampling function $f(i, K)$ then draws K support indices from V using the restricted and renormalized probability distribution over candidates:

$$f(i, K) = \begin{cases} \text{Sample}(V, K, \mathbf{p}_V^{(t)}), & \text{if } |V| \geq K \\ \emptyset, & \text{if } |V| < K \end{cases} \quad (8)$$

where $\mathbf{p}_V^{(t)}$ is the renormalized distribution restricted to V . This strategy ensures that CoTs which have consistently acted as useful support examples for improving other reasoning chains are more frequently sampled, allowing the model to prioritize high-quality and influential reasoning structures during refinement.

B Dataset Statistics

The dataset is annotated for persuasive spans for 20 techniques. The corpus contains 3922 annotated spans, out of which 3126 are allocated to the train set and 796 to the test set.

B.1 Technique-wise Distribution

The dataset identifies 20 specific propaganda techniques. The distribution across these techniques is characterized by significant variance, reflecting the real-world prevalence of certain rhetorical strategies over others.

High-Frequency Techniques The most prevalent technique across both datasets is Smears (T15), totaling 851 spans. This is followed by Assertion (T19, 561 spans) and Appeal to emotion/fear/prejudice (T9, 515 spans). The prominence of these techniques indicates a heavy reliance on ad hominem attacks and emotional manipulation within the sampled discourse.

Phase	Configuration	Mistral (s)	Qwen (s)
Inference	$R = 3$	6.05	11.70
Refinement	$K = 1$	2.99	5.63
	$K = 2$	4.70	9.01
	$K = 3$	6.12	12.00

Table 8: Inference and refinement time (in seconds) for Mistral and Qwen.

Low-Frequency Techniques Conversely, techniques such as Reductio ad Hitlerum (T5) and Black and White Fallacy (T2) show significantly lower frequencies (28 and 32 total spans, respectively).

C Prompt

During iterative refinement, we employed a fixed Chain-of-Thought prompt that requires step-by-step reasoning for a target propaganda technique before any span extraction. Spans are produced only upon a positive reasoning conclusion based on F_{LCS} scores, ensuring interpretable and evaluation-consistent refinement. The prompts used during refinement and inference are included in Figure 6.

D Global pool convergence over PTC

Figures 4 illustrate Global pool convergence curves over the PTC dataset. We observe a similar convergence trend for PTC as observed for INDI-PROPspan, with a slight difference during the initial phase. While INDI-PROPspan showed a steep performance increase early in the refinement, PTC shows a more gradual increase through the initial phase. The update density convergence trend is similar to the INDI-PROPspan dataset, with large updates in the initial phase, then steady near zero degrees by the final iteration. INDI-PROPspan is an event-specific dataset, allowing in-context examples from the same event to yield rapid performance gains, whereas PTC is not event-specific, leading to a steadier convergence trajectory. Overall, both curves corresponding to both datasets demonstrate convergence.

E Inter-Bias Inference analysis

Figure 5 compares the techniques across all three bias strategies, showing that persuasive techniques exhibit the same trends for in-context example size during inference, indicating similar intrinsic cues acquired during the refinement process, with sampling strategy being i-bias. However, for j and random bias, this trend shows slight variations across

frequent and infrequent techniques, such as doubt and flag-waving, which are less dominant. The trajectory of F_{LCS} vs R shows a difference, indicating that the refinement curve flattens but not towards meaningful intrinsic cues, as j-bias prioritises the CoTs that cause updates, resulting in an echo chamber of cues that were successful in earlier refinements but failed at inference. Similarly, random, as a stochastic process, was unable to capture meaningful reasons either, whereas dominant techniques, name-calling, and loaded language performed slightly better during refinement, due to the large population size.

F Result Analysis

Table 13 reports of R_{LCS} , P_{LCS} , and $F1_{LCS}$ across all baselines and ReCoT configurations at $K=1$, $K=2$ and $K=3$ on both INDI-PROPspan and PTC, extending the main $K=2$ results in Table 2.

Mistral at $K=1$, $R=3$ achieves $F1_{LCS}=0.354$ and 0.360 for macro and weighted respectively and $R_{LCS}=0.447$ a substantial recall gain over zero-shot ($R_{LCS}=0.270$) and 1-shot ($R_{LCS}=0.365$), indicating that even a single in-context example from the refined pool is sufficient to improve span coverage. Qwen at $K=1$, $R=3$ achieves $F1_{LCS}=0.321$ and the highest weighted P_{LCS} of 0.380 among $K=1$ configurations, suggesting that recot at this setup accurately captures span boundaries. GPT-4o-mini at $K=1$ shows the weakest performance across all R values, achieving the highest $F1_{LCS}=0.307$ at $R=4$ for macro as well as weighted and degrades at $R=5$, still surpassing its own zero-shot baseline (0.294) and CoT baseline (0.302), confirming that population-level refinement provides a consistent benefit regardless of backbone capability.

On PTC dataset at $K=1$, Mistral achieves its best result at $R=4$ achieves $F1_{LCS}=0.193$ and 0.205 for macro and weighted, respectively, with Qwen performing comparably at $R=1$ and $R=2$ ($F1_{LCS}$ for W.Avg = $0.201 - 0.205$), both surpassing the Mistral prompting baselines on PTC. At $K=3$, performance degrades relative to $K=1$ and $K=2$ configurations for all LLM backbones, with the weakest observed performance is Qwen, where $F1_{LCS}$ drops to 0.194 to as low as 0.166 on PTC dataset, falling below zero-shot performance, this degradation at $K=3$ is also observed for macro recall and precision on INDI-PROPspan, Mistral at $K=3$ maintains relatively high recall at $R=5$ ($R_{LCS}=0.437-0.478$) but shows markedly reduced precision ($P_{LCS}=0.308-$

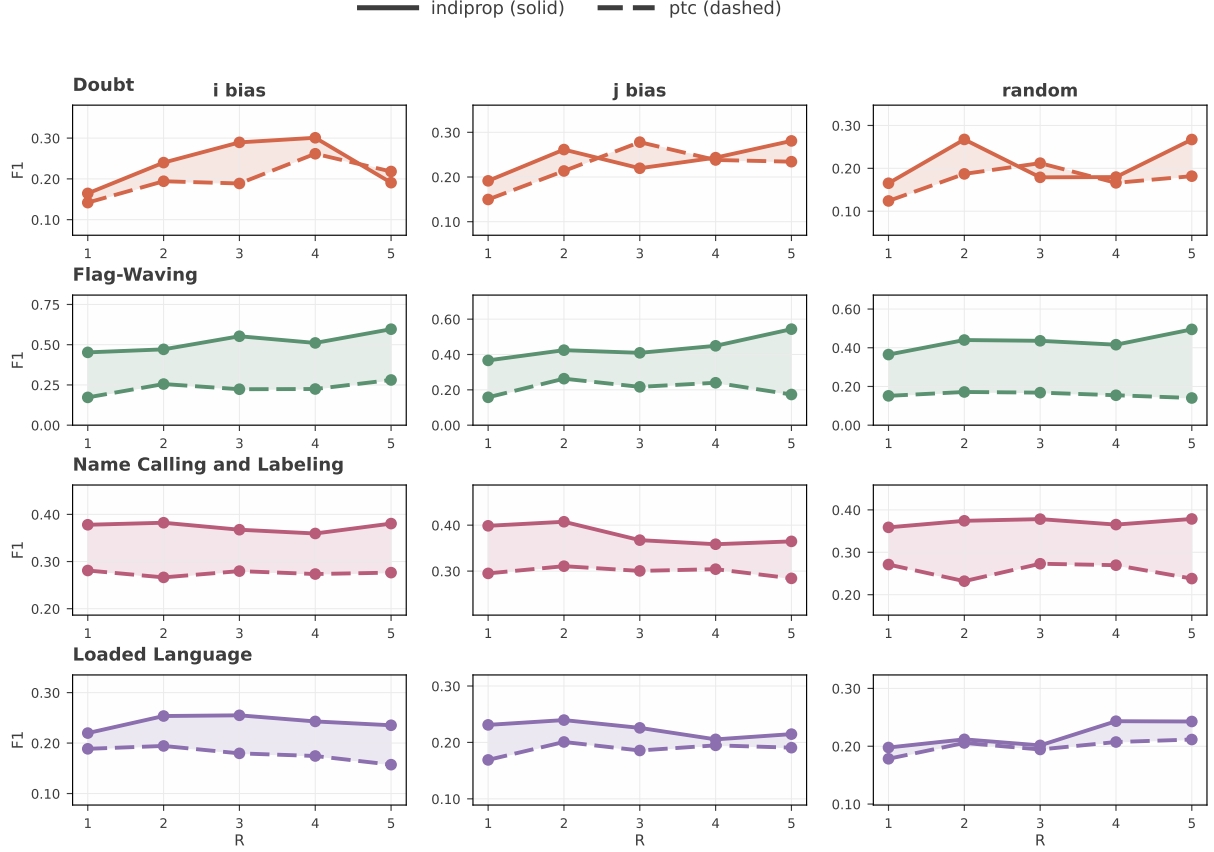


Figure 5: Inter-bias comparison of technique wise inference results across R values for $K = 2$ on Mistral backbone

0.331) compared to baselines, indicating over-conditioning due to the presence of three in-context examples causes the model to over-predict spans, expanding coverage at the cost of boundary precision.

For overall recall performance, GPT EvoPrompt achieves the highest recall on the INDI-PROPspan dataset among all methods. But for the same precision, it is as low as 0.237-0.262, the lowest among all GPT configurations. The $F1_{LCS}$ collapses to 0.272-0.300. This behaviour indicates EvoPrompt’s optimisation drives the model toward over-prediction, generating excessively long or numerous spans that happen to overlap with ground truth. Similar behaviour is noticed on the PTC dataset, where recall reaches to 0.419 Avg and 0.366 W.Avg, but precision drops to 0.200-0.178, again indicating over-prediction.

The performance of ReCoT at $K=2$ in terms of precision and recall. Mistral at $R=3$ achieves the best ReCoT R_{LCS} of 0.457 for macro Avg and 0.446 for Weighted Avg. on INDI-PROPspan dataset. Beyond $R=3$, recall drops slightly ($R=4$: 0.420, $R=5$: 0.429), indicating that $R=3$ strikes a

balance between constrained and noisy contexts. Qwen achieves best macro P_{LCS} at $R=5$: 0.368 and weighted P_{LCS} at $R=3$: 0.387. This behaviour indicates that Mistral’s refined CoTs are coverage-oriented, whereas Qwen’s refined CoTs are more boundary-precise, making fewer but more accurate predictions. Both strategies eventually converge to competitive F1 along different paths, indicating that the global pool converges in the end. EvoPrompt(GPT) achieved 0.419 macro R_{LCS} , 0.366 weighted R_{LCS} on the PTC dataset. As stated earlier, these are the result of overprediction. And a similar trend is observed for EvoPrompt on INDI-PROPspan.

Qwen achieves the best ReCoT recall and the best ReCoT precision on PTC dataset simultaneously at $R=4$: macro R_{LCS} of 0.256 and weighted macro R_{LCS} of 0.250, which is why it also produces the best $F1_{LCS}$ of 0.200-0.211 (mentioned in Table 2). Whereas GPT achieves macro recall of 0.342–0.362 on INDI-PROPspan dataset, comparable to Qwen, but its weighted recall (0.225–0.248) is dramatically lower. Overall, these results confirm that ReCoT recall improvements are grounded in

technique understanding rather than span inflation and achieve competitive precision alongside recall gains across backbones and datasets. This case provides evidence that population-level CoT refinement captures the intrinsic definition characteristic of the techniques

Config	Backbone	IndiProp		PTC	
		Avg	W.Avg	Avg	W.Avg
0-shot	GPT	0.231	0.212	0.114	0.170
1-shot		0.176	0.201	0.170	0.203
CoT		0.241	0.244	0.152	0.194
RePrompt	GPT	0.106	0.158	0.140	0.184
ReCoT R1		0.229	0.246	0.100	0.141
ReCoT R2		0.235	0.251	0.102	0.141
ReCoT R3		0.253	0.246	0.092	0.148
ReCoT R4		0.264	0.252	0.130	0.161
ReCoT R5		0.272	0.274	0.098	0.147

Table 9: Performance comparison of the GPT backbone on the SemEval benchmark. Average F1 (Avg) and Weighted Average F1 (W.Avg) are reported for the IndiProp and PTC tasks.

Inference Time: The average inference time per instance for the $R = 3$ configuration for INDI-PROPspan during inference is approximately 6.05s for Mistral and 11.70s for Qwen. And, for the refinement phase, the runtime for a single instance within a single iteration is approximately 2.99s (Mistral) and 5.63s (Qwen) for K=1, 4.7s (Mistral) and 9.01s (Qwen) for K=2, and 6.12s (Mistral) and 12.0s (Qwen) for K=3. (as shown in table 8)

Technique Wise Performance: Tables 10 and 12 report the per-technique F1 scores for the *Mistral K2 - iBias* configuration on the INDI-PROPspan and PTC datasets, respectively, when varying the number of in-context examples used during inference. Here, R1-R5 denote inference with one to five retrieved Chain-of-Thought (CoT) examples from the refined example pool for the corresponding propaganda technique. While the main paper reports aggregate performance, these tables provide a fine-grained view of how increasing the amount of retrieved in-context evidence affects individual techniques.

Across both datasets, the influence of additional in-context examples is technique-dependent rather than monotonic. Some techniques, such as *Flag Waving*, benefit from additional examples, whereas others, including *Loaded Language* and *Name Calling and Labelling*, remain comparatively stable. Moreover, the same technique does not necessarily exhibit similar trends across datasets; for in-

Technique	R1	R2	R3	R4	R5
Causal Oversimplification	0.308	0.291	0.258	0.208	0.256
Black-and-white Fallacy	0.347	0.384	0.327	0.371	0.356
Straw Man	0.280	0.183	0.269	0.160	0.291
Whataboutism	0.309	0.302	0.375	0.331	0.340
Reductio ad Hitlerum	0.299	0.577	0.625	0.465	0.625
Red Herring	0.346	0.483	0.377	0.523	0.448
Loaded Language	0.220	0.254	0.255	0.243	0.235
Name Calling and Labeling	0.378	0.382	0.368	0.359	0.381
Appeal to emotion/fear/prejudice	0.335	0.360	0.382	0.409	0.378
Slogans	0.484	0.486	0.495	0.412	0.527
Flag-Waving	0.452	0.471	0.553	0.511	0.596
Exaggeration and Minimization	0.225	0.245	0.168	0.223	0.286
Thought-terminating Cliches	0.257	0.319	0.173	0.216	0.234
Bandwagon	0.554	0.369	0.551	0.390	0.284
Smears	0.356	0.331	0.386	0.391	0.334
OVC	0.202	0.276	0.290	0.452	0.412
Doubt	0.164	0.240	0.290	0.301	0.191
Appeal to Authority	0.297	0.238	0.418	0.302	0.332
Assertion	0.381	0.413	0.399	0.386	0.421
Glittering Generalities (Virtue)	0.309	0.316	0.373	0.308	0.320

Table 10: IndiProp per-technique F1 scores across refinement rounds (Mistral - K2(i-bias)).

stance, *Bandwagon* shows contrasting behaviour on IndiProp and PTC. These observations suggest that the optimal amount of retrieved CoT evidence depends on both the propaganda technique and the dataset, reflecting differences in class distributions and the linguistic characteristics of the two corpora.

Bias strategies: In addition to the proposed selection strategy, we evaluate different bias strategies to examine the effects of exploratory and exploitative biases during sampling and in the refinement process. **Random sampling:** This sampling strategy is randomised, that is, it uniformly samples without any preference from the valid set of indices explained in Section 4. **J-bias:** It increases the probability of CoTs that have influenced and improved other reasoning chains. Table 11 compares I-Bias, Random, and J-Bias across $R \in \{1, 2, 3, 4, 5\}$ with K=2 on Mistral and Qwen inference backbones. Random at R=5 achieved the highest macro $F1_{LCS}$ of 0.378 on Mistral backbone for the INDI-PROPspan dataset, but the same configuration yields the lowest weighted $F1_{LCS}$ of 0.111 on PTC. As random is a stochastic process, consistent behaviour cannot be guaranteed. J-Bias, exclusively prioritising CoTs that have previously updated other CoTs, risks creating an echo chamber of dominant reasoning cues that were successful in the early refinement phase. I-Bias, by contrast, balances exploration and exploitation by rewarding CoTs based on their own performance improvement while keeping the bias increment small to retain diversity.

SemEval Span-overlap Metrics: We reran the evaluation using SemEval’s official span-overlap metrics, as shown in Table 9. On IndiProp, Re-

Setup	R	Mistral				Qwen			
		IndiProp		PTC		IndiProp		PTC	
		Avg	W.Avg	Avg	W.Avg	Avg	W.Avg	Avg	W.Avg
I - BIAS	R=1	0.320	0.311	0.167	0.180	0.296	0.326	0.172	0.191
	R=2	0.343	0.300	0.166	0.182	0.316	0.348	0.181	0.196
	R=3	0.364	0.347	0.155	0.172	0.345	0.362	0.184	0.198
	R=4	0.344	0.317	0.172	0.170	0.326	0.356	0.200	0.211
	R=5	0.361	0.306	0.172	0.170	0.359	0.355	0.170	0.187
RANDOM	R=1	0.328	0.355	0.175	0.120	0.304	0.327	0.186	0.137
	R=2	0.343	0.348	0.169	0.122	0.326	0.344	0.182	0.137
	R=3	0.367	0.377	0.168	0.126	0.347	0.362	0.172	0.124
	R=4	0.360	0.360	0.162	0.117	0.333	0.350	0.181	0.135
	R=5	0.378	0.361	0.153	0.111	0.328	0.351	0.178	0.134
J - BIAS	R=1	0.333	0.362	0.166	0.160	0.315	0.337	0.181	0.173
	R=2	0.352	0.366	0.188	0.184	0.306	0.334	0.184	0.174
	R=3	0.343	0.367	0.171	0.173	0.319	0.344	0.170	0.177
	R=4	0.331	0.371	0.174	0.174	0.326	0.343	0.177	0.178
	R=5	0.360	0.362	0.161	0.161	0.328	0.350	0.176	0.168

Table 11: ReCoT performance across bias strategies with $K = 2$ configuration.

Technique	R1	R2	R3	R4	R5
Appeal to Authority	0.128	0.137	0.147	0.153	0.128
Appeal to fear/prejudice	0.176	0.211	0.173	0.195	0.202
Bandwagon, Reductio ad Hitlerum	0.253	0.167	0.167	0.167	0.167
Black-and-white Fallacy	0.047	0.106	0.089	0.059	0.084
Causal Oversimplification	0.182	0.157	0.111	0.224	0.197
Doubt	0.142	0.194	0.189	0.262	0.218
Exaggeration and Minimization	0.133	0.086	0.087	0.097	0.097
Flag Waving	0.173	0.256	0.223	0.225	0.281
Loaded Language	0.189	0.194	0.180	0.174	0.157
Name Calling and Labeling	0.281	0.266	0.280	0.274	0.276
Repetition	0.172	0.116	0.108	0.105	0.110
Slogans	0.133	0.147	0.125	0.130	0.133
Thought-terminating Cliches	0.136	0.074	0.075	0.053	0.126
Whataboutism, Strawmen and Red Herring	0.197	0.209	0.219	0.221	0.174

Table 12: PTC per-technique F1 scores across refinement rounds (Mistral - k2(i-bias)).

CoT outperforms both SFT and standard prompting baselines on both backbones, peaking at avg. F1 0.272 / w.F1 0.274 (GPT R5). On PTC, scores are lower overall but follow the same trend, with GPT ReCoT peaking at R4 (avg. F1 0.130, w.F1 0.161), close to the strongest prompting baseline (GPT 1-shot, w.F1 0.203). These results mirror the trend we reported with our LCS-based metric.

G Before-and-After examples

Table 14 compared iteration 0 with iteration 25 across all bias strategies. Compared to the zero-shot initialization, all refined variants produce more focused reasoning. I-Bias learns to directly evaluate whether the article misrepresents the opponent’s viewpoint, resulting in a concise CoT that identifies the key propagandistic span and achieves the highest F1 (0.706). J-Bias improves over initialization by providing a more structured explanation that better captures multiple propaganda cues, although its broader reasoning extracts an additional, less relevant span, leading to a lower F1 than I-Bias. Random refinement is also more targeted than the initial zero-shot CoT, but its inconsistent explo-

ration identifies only a single partial span, yielding only a modest improvement over initialization. Overall, iterative refinement makes the reasoning progressively more focused and evidence-driven compared to the unrefined zero-shot CoT.

Setup	Backbone	Config	IndiProp						PTC					
			R_{LCS}		P_{LCS}		$F1_{LCS}$		R_{LCS}		P_{LCS}		$F1_{LCS}$	
			Avg	W.Avg	Avg	W.Avg	Avg	W.Avg	Avg	W.Avg	Avg	W.Avg	Avg	W.Avg
SFT	Hitachi Aschern	–	0.381	0.438	0.150	0.201	0.180	0.232	0.179	0.217	0.084	0.108	0.093	0.118
		–	0.311	0.345	0.203	0.236	0.205	0.231	0.248	0.296	0.206	0.219	0.174	0.203
Prompting	Mistral	0-shot	0.270	0.276	0.319	0.345	0.263	0.273	0.259	0.243	0.199	0.211	0.187	0.184
		1-shot	0.365	0.346	0.340	0.365	0.318	0.317	0.244	0.215	0.214	0.225	0.200	0.192
		2-shot	0.154	0.157	0.154	0.157	0.154	0.157	0.250	0.225	0.218	0.237	0.195	0.194
		CoT	0.343	0.352	0.294	0.320	0.285	0.300	0.244	0.253	0.187	0.210	0.186	0.202
Reasoning	Mistral	RePrompting	0.382	0.369	0.352	0.349	0.326	0.314	0.258	0.258	0.158	0.197	0.160	0.186
		EvoPrompt	0.256	0.306	0.217	0.227	0.213	0.227	0.202	0.242	0.182	0.204	0.171	0.195
Prompting	GPT	0-shot	0.314	0.285	0.345	0.355	0.294	0.279	0.194	0.173	0.197	0.204	0.169	0.163
		1-shot	0.249	0.253	0.320	0.356	0.254	0.263	0.259	0.258	0.237	0.239	0.218	0.215
		2-shot	0.269	0.267	0.341	0.371	0.275	0.279	0.278	0.261	0.226	0.223	0.212	0.203
		CoT	0.313	0.297	0.361	0.364	0.302	0.289	0.265	0.257	0.234	0.231	0.210	0.208
Reasoning	GPT	RePrompting	0.306	0.312	0.287	0.291	0.297	0.308	0.228	0.234	0.216	0.246	0.193	0.209
		EvoPrompt	0.484	0.515	0.237	0.262	0.272	0.300	0.419	0.366	0.200	0.178	0.232	0.208
$K = 1$	Mistral	ReCoT R1	0.381	0.377	0.330	0.367	0.323	0.336	0.227	0.211	0.185	0.220	0.171	0.185
		ReCoT R2	0.400	0.390	0.348	0.360	0.340	0.337	0.232	0.225	0.184	0.218	0.171	0.185
		ReCoT R3	0.447	0.447	0.349	0.370	0.354	0.360	0.234	0.229	0.200	0.236	0.182	0.196
		ReCoT R4	0.402	0.413	0.337	0.362	0.334	0.345	0.242	0.236	0.211	0.245	0.193	0.205
		ReCoT R5	0.424	0.427	0.364	0.379	0.357	0.360	0.236	0.229	0.200	0.230	0.186	0.196
	Qwen	ReCoT R1	0.356	0.366	0.340	0.371	0.323	0.339	0.233	0.230	0.194	0.228	0.184	0.201
		ReCoT R2	0.338	0.372	0.335	0.376	0.313	0.344	0.230	0.235	0.201	0.230	0.190	0.205
		ReCoT R3	0.368	0.398	0.331	0.380	0.321	0.358	0.230	0.234	0.199	0.222	0.187	0.199
		ReCoT R4	0.385	0.390	0.351	0.375	0.341	0.352	0.217	0.222	0.192	0.222	0.177	0.192
		ReCoT R5	0.368	0.377	0.340	0.369	0.323	0.342	0.211	0.220	0.182	0.207	0.170	0.185
	GPT	ReCoT R1	0.313	0.299	0.333	0.345	0.296	0.289	0.201	0.207	0.209	0.238	0.178	0.194
		ReCoT R2	0.315	0.326	0.323	0.351	0.290	0.304	0.204	0.218	0.185	0.215	0.169	0.190
		ReCoT R3	0.312	0.322	0.320	0.333	0.287	0.293	0.202	0.217	0.172	0.199	0.166	0.185
		ReCoT R4	0.330	0.333	0.332	0.345	0.307	0.307	0.210	0.220	0.185	0.215	0.172	0.191
		ReCoT R5	0.328	0.324	0.326	0.337	0.301	0.299	0.207	0.218	0.184	0.217	0.171	0.192
$K = 2$	Mistral	ReCoT R1	0.380	0.351	0.320	0.321	0.320	0.311	0.241	0.223	0.170	0.202	0.167	0.180
		ReCoT R2	0.404	0.343	0.349	0.311	0.343	0.300	0.239	0.227	0.171	0.206	0.166	0.182
		ReCoT R3	0.457	0.446	0.354	0.330	0.364	0.347	0.224	0.216	0.161	0.194	0.155	0.172
		ReCoT R4	0.420	0.375	0.334	0.318	0.344	0.317	0.239	0.216	0.170	0.183	0.172	0.170
		ReCoT R5	0.429	0.364	0.357	0.303	0.361	0.306	0.243	0.215	0.172	0.183	0.172	0.170
	Qwen	ReCoT R1	0.338	0.357	0.311	0.353	0.296	0.326	0.210	0.216	0.180	0.217	0.172	0.191
		ReCoT R2	0.345	0.375	0.334	0.380	0.316	0.348	0.222	0.225	0.191	0.221	0.181	0.196
		ReCoT R3	0.388	0.401	0.352	0.387	0.345	0.362	0.233	0.233	0.194	0.222	0.184	0.198
		ReCoT R4	0.363	0.393	0.343	0.379	0.326	0.356	0.256	0.250	0.209	0.233	0.200	0.211
		ReCoT R5	0.406	0.396	0.368	0.377	0.359	0.355	0.226	0.225	0.175	0.204	0.170	0.187
	GPT	ReCoT R1	0.342	0.225	0.347	0.243	0.322	0.215	0.218	0.151	0.184	0.144	0.177	0.128
		ReCoT R2	0.325	0.231	0.321	0.236	0.297	0.214	0.196	0.138	0.187	0.149	0.168	0.124
		ReCoT R3	0.358	0.248	0.336	0.229	0.320	0.216	0.212	0.154	0.175	0.147	0.165	0.130
		ReCoT R4	0.362	0.248	0.318	0.224	0.313	0.215	0.202	0.145	0.181	0.152	0.164	0.128
		ReCoT R5	0.341	0.245	0.314	0.228	0.301	0.215	0.183	0.138	0.175	0.148	0.160	0.127
$K = 3$	Mistral	ReCoT R1	0.436	0.410	0.341	0.340	0.347	0.333	0.231	0.193	0.155	0.175	0.156	0.153
		ReCoT R2	0.394	0.412	0.316	0.346	0.319	0.340	0.218	0.195	0.162	0.177	0.157	0.157
		ReCoT R3	0.438	0.478	0.325	0.323	0.334	0.338	0.237	0.210	0.166	0.176	0.168	0.165
		ReCoT R4	0.432	0.479	0.307	0.309	0.324	0.331	0.230	0.199	0.152	0.167	0.153	0.152
		ReCoT R5	0.437	0.478	0.308	0.331	0.328	0.347	0.226	0.201	0.152	0.167	0.156	0.156
	Qwen	ReCoT R1	0.371	0.382	0.345	0.367	0.330	0.343	0.232	0.229	0.208	0.233	0.194	0.204
		ReCoT R2	0.345	0.375	0.332	0.373	0.310	0.343	0.224	0.229	0.197	0.229	0.183	0.199
		ReCoT R3	0.396	0.419	0.345	0.364	0.340	0.355	0.218	0.219	0.189	0.215	0.178	0.192
		ReCoT R4	0.387	0.418	0.331	0.347	0.326	0.343	0.209	0.213	0.174	0.203	0.166	0.182
		ReCoT R5	0.410	0.425	0.353	0.365	0.343	0.353	0.227	0.217	0.184	0.198	0.180	0.183
	GPT	ReCoT R1	0.347	0.186	0.312	0.180	0.305	0.168	0.204	0.179	0.193	0.200	0.175	0.166
		ReCoT R2	0.328	0.187	0.297	0.168	0.285	0.162	0.188	0.177	0.171	0.186	0.155	0.158
		ReCoT R3	0.341	0.194	0.290	0.168	0.286	0.162	0.188	0.179	0.159	0.174	0.149	0.153
		ReCoT R4	0.334	0.191	0.295	0.170	0.287	0.163	0.195	0.183	0.169	0.183	0.157	0.159
		ReCoT R5	0.348	0.194	0.303	0.174	0.298	0.168	0.196	0.187	0.161	0.177	0.154	0.159

Table 13: $F1_{LCS}$, R_{LCS} , and P_{LCS} for ReCoT with I_{bias} at $K \in \{1, 2, 3\}$ across Mistral, Qwen, and GPT on INDI-PROPSpan and PTC

Variant	Evaluation Details
Initialization ($n = 0$)	<CoT> Step 1: Identify the main arguments presented in the article. The article discusses the contrasting actions of citizens during a public appreciation event for frontline workers and the protests against CAA, NRC, and NPR led by Suranya Aiyar. Step 2: Look for instances where an opposing viewpoint is misrepresented. The article portrays Suranya Aiyar’s protests as being out of touch or inappropriate given the context of a pandemic and the collective appreciation of healthcare workers, suggesting that those who protest are disregarding the situation at hand. Step 3: Analyze specific statements that may reduce the opposing viewpoint to an unfairly simple or easily attackable position. The article claims "partisan political interests even in the face of an apocalypse." This implies that any dissent against the government’s practices during a crisis is wrong or irrelevant, which misrepresents the nuanced argument that some may hold against these policies, portraying them as entirely neglectful of the situation. Step 4: The statement about residents of Jamia Nagar raising slogans while others rallied for healthcare workers further simplifies their stance to mere opposition, not recognizing the validity of their protest or their right to express it. </CoT> Extracted Span(s): “some people would not abandon their partisan political interests even in the face of an apocalypse.” “On a similar note, when the nation demonstrated their appreciation towards those providing people with healthcare in times of this major crisis, residents of Jamia Nagar’s Abul Fazal colony raised ‘Azaadi’ slogans instead.” F1 Score: 0.0841
I-Bias ($n = 25$)	<CoT> Step 1: The article describes the actions of Suranya Aiyar and others during a nationwide “Janta Curfew” to show appreciation for frontline workers battling the pandemic. Step 2: Suranya Aiyar is portrayed as screaming slogans demanding the rollback of CAA, NRC, and NPR, which are contentious issues in India. Step 3: The article criticizes her actions by stating that she “riled against the people as the nation came together as one,” implying that her actions were disruptive and inappropriate during a time of national unity. Step 4: To identify Straw Man propaganda, we need to check if the article misrepresents or exaggerates Suranya Aiyar’s actions or intentions. The statement “riled against the people as the nation came together as one” could be interpreted as a Straw Man if it attributes an extreme or simplistic viewpoint to her actions that may not be accurate. Conclusion: Propaganda spans exist. </CoT> riled against the people as the nation came together as one Extracted Span(s): “riled against the people as the nation came together as one” F1 Score: 0.7060
J-Bias ($n = 25$)	<CoT> Step 1: To determine if there are any “Straw Man” arguments in the article, I need to identify whether the article misrepresents or oversimplifies the position of any opposing side or group. A Straw Man argument often involves refuting a weaker version of the actual argument presented by the opposing side. Step 2: Analyzing the article, it describes the actions of Suranya Aiyar and other individuals who raised slogans during the ‘Janta Curfew’. The article portrays these actions as disruptive and partisan, contrasting them with the nationwide effort to show appreciation for frontline workers battling the pandemic. Step 3: I look for any indication that the article misrepresents the motivations or arguments of Suranya Aiyar or the other individuals mentioned. The article states that Suranya Aiyar and others “riled against the people as the nation came together as one,” and that their actions demonstrate a refusal to abandon “partisan political interests even in the face of an apocalypse.” This could be seen as a misrepresentation or oversimplification of their actions, as it frames their political statements as inherently disruptive and selfish, without exploring their underlying motivations or the context of their slogans. Conclusion: Propaganda spans exist. </CoT> riled against the people as the nation came together as one some people would not abandon their partisan political interests even in the face of an apocalypse Extracted Span(s): “riled against the people as the nation came together as one” “some people would not abandon their partisan political interests even in the face of an apocalypse” F1 Score: 0.5500
Random ($n = 25$)	<CoT> Step 1: To identify if there are any statements that qualify as straw man propaganda, I need to check if the article misrepresents an opponent’s position in order to make it easier to attack or refute. Step 2: The article discusses Suranya Aiyar’s actions and slogans during the ‘Janta Curfew’. It mentions her demands for the rollback of CAA, NRC, and NPR, and her slogans in support of various causes. Step 3: The article states, “Suranya Aiyar was walking around her lawn presumably while she raised these slogans at 5 pm on Sunday.” This could be interpreted as a misrepresentation if the context or intent behind her actions is not accurately portrayed. Additionally, the phrase “some people would not abandon their partisan political interests even in the face of an apocalypse” could be seen as a straw man argument if it oversimplifies or misrepresents the motivations of those opposing CAA, NRC, and NPR. Conclusion: Propaganda spans exist. </CoT> some people would not abandon their partisan political interests even in the face of an apocalypse. Extracted Span(s): “some people would not abandon their partisan political interests even in the face of an apocalypse.” F1 Score: 0.1000

Table 14: Propaganda Span and Reasoning (CoT) Results

Prompt Used During Iterative Refinement
You are an expert propaganda span detector. Your task is to carefully analyze the article and decide step by step (Chain of Thought) whether there are any statements that qualify as propaganda of type { technique}. Instructions: 1. Use a step-by-step reasoning process inside <CoT> . . . </CoT> tags. 2. If no spans are found, clearly state that in the CoT and output nothing inside tags. 3. If spans are found, wrap each exact substring from the article in . . . tags. Output format (exactly): <pre> <CoT> Step 1: ... Step 2: ... Step 3: ... Conclusion: Propaganda spans exist / do not exist </CoT> <!-- zero or more spans, depending on conclusion --> </pre>
Prompt Used During Inference
Here is an example: {support text} Target Article: {full article} You are an expert propaganda span detector. Your task is to carefully analyze the article and decide step by step (Chain of Thought) whether there are any statements that qualify as propaganda of type { technique}. Instructions: 1. Use a step-by-step reasoning process inside <CoT> . . . </CoT> tags to decide if such propaganda spans exist. 2. If no spans are found, clearly state that in the CoT and output nothing inside tags. 3. If spans are found, wrap each exact substring from the article in . . . tags. Output format (exactly): <pre> <CoT> Step 1: ... Step 2: ... Step 3: ... Conclusion: Propaganda spans exist/do not exist </CoT> <!-- zero or more spans, depending on conclusion --> </pre>

Figure 6: Prompts used in our framework: (top) iterative refinement prompt, (bottom) inference prompt.