Fine-tuning with Hierarchical Prompting for Robust Propaganda Classification Across Annotation Schemas

Lukas Stähelin[1]*, Veronika Solopova[1,2]*, Max Upravitelev[1,2], David Kaplan[1], Ariana Sahitaj[1,2]

Premtim Sahitaj[1,2], Charlott Jakob[1,2], Sebastian Möller[1,2], Vera Schmitt[1,2,3]

[1]Technische Universität Berlin, QU Lab, XplaiNLP Group, Berlin, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
[3]Centre for European Research in Trusted AI (CERTAIN)

Correspondence: veronika.solopova@tu-berlin.de

Abstract

Propaganda detection in social media is challenging due to noisy, short texts and low annotation agreements. We introduce a new intent-focused taxonomy of propaganda techniques and compare it against an established, higher-agreement schema. Along three dimensions (model portfolio, schema effects, and prompting strategy) we evaluate the taxonomies as a classification task with the help of four language models (GPT-4.1-nano, Phi-4 14B, Qwen2.5-14B, Qwen3-14B). Our results show that fine-tuning is essential, since it transforms weak zero-shot baselines into competitive systems and reveals methodological differences that are hidden using base models. Across schemas, the Qwen models achieve the strongest overall performance, and Phi-4 14B consistently outperforms GPT-4.1-nano. Our hierarchical prompting method (**HiPP**), which predicts fine-grained techniques before aggregating them, is especially beneficial after fine-tuning and on the more ambiguous, low-agreement taxonomy, while remaining competitive on the simpler schema. The HQP dataset, annotated with the new intent-based labels, provides a richer lens on propaganda's strategic goals and a challenging benchmark for future work on robust, real-world detection.

1 Introduction

Propaganda has long been used to shape public opinion and influence political outcomes (Jowett and O'donnell, 2018), while online platforms amplify its reach. The Russia–Ukraine war exemplifies this evolution, with propaganda deployed as part of hybrid warfare (Perez, 2022; Zhdanova and Orlova, 2017). Russian campaigns manipulate perceptions, sow distrust, and polarize societies, making automatic detection of propaganda vital for safeguarding democratic processes (Wardle and Derakhshan, 2017; Bayer et al., 2021).

While propaganda detection has been explored in long-form news articles (Martino et al., 2020), short-form social media introduces additional challenges: limited context, informal language, and heavy use of abbreviations (Vijayaraghavan and Vosoughi, 2022; Maarouf et al., 2024). Subtle linguistic cues and context-dependent meanings often lead to low inter-annotator agreement (Srba et al., 2024; Hasanain et al., 2024). Propaganda also evolves over time (Solopova et al., 2023, 2024), with AI-generated content becoming increasingly prevalent on social media (Solopova et al., 2025). These challenges are amplified by trade-offs in annotation schema design. Realistic schemas capture the diversity of real-world propaganda but reduce annotation reliability, while simplified ones boost agreement and learnability at the cost of ecological realism. Understanding how methodological choices interact with schema design is essential for building robust detection systems. Existing taxonomies, such as the span-level schemas by Da San Martino et al. (2020); Sprenkamp et al. (2023) and the recent hybrid hierarchy of Sahitaj et al. (2025), are all technique-centric, enumerating persuasive rhetorical devices (e.g., loaded language, appeals to fear). Our new schema instead groups surface techniques by communicative intent, introducing high-level categories such as *Shift Blame and Justify Aggression*, *Manufacture Consent and Identity*, and *Confuse and Distract*. These labels capture strategic goals like reframing moral responsibility, building in-group loyalty, or generating uncertainty. Prior schemas either scatter these constructs across multiple tags or do not explicitly encode. By reorganising techniques around these intents, our taxonomy is conceptually closer to framing-style analyses and provides a more explanatory lens on why a message is propagandistic, although its empirical advantages beyond the current dataset remain to be validated.

In addition, we propose a framework for ana-

*Equal contribution

Findings of the Association for Computational Linguistics: ACL 2026, pages 2887–2902

lyzing LLM-based propaganda classification along three interacting dimensions: (i) annotation schema design and its associated reliability, (ii) supervision regime (zero-shot vs. fine-tuned), and (iii) prompting strategy (direct vs. hierarchical). Importantly, we treat these three dimensions as orthogonal factors. The annotation schema defines the label space and its reliability (e.g., inter-annotator agreement), the supervision regime determines whether models operate in zero-shot or are adapted via fine-tuning, and the prompting strategy specifies how predictions are structured (direct vs. hierarchical decomposition). In particular, the annotation schema should not be conflated with hierarchical prompting: while both introduce structure, the former concerns label design, whereas the latter defines the inference procedure applied to a fixed label space. Prior work typically varies these factors independently or implicitly. By controlling supervision and holding label information constant, we isolate the effect of hierarchical decomposition and show how its benefits depend on annotation noise and model adaptation. Our contributions are primarily methodological and threefold:

1. We introduce an intent-based taxonomy of propaganda techniques, the first entirely new conceptual taxonomy since 2019;

2. We conduct a controlled comparison of our custom schema and the recent taxonomy from Sahitaj et al. (2025) and error analysis of results. For this, we also create high level labels, using a new clustering procedure, resulting in the highest IAA in human validation for propaganda annotation;

3. For both schemas, we compare 4 models (Phi-4, GPT-4.1-nano, Qwen2.5 and 3) and two approaches to high-level classification: **Direct-High**, where models predict high-level categories directly, and the hierarchical **Main→High** strategy, where fine-grained labels are predicted first and then aggregated.

Our results and through error analysis show that fine-tuning is decisive for improving performance across models. We also demonstrate that our hierarchical propaganda prompting (**HiPP**) method improves high-level classification for our schema once models are fine-tuned, offering a stable methodological advantage in settings with noisy and imbalanced annotations. Overall, the contribution lies in analysing how different factors interact. Our data, labels and code are available in a Github[1]. Best performing fine-tuned models are accessible in a Hugging face repository[2].

2 Related Work

Datasets. Da San Martino et al. (2019) introduced the first dataset for fine-grained propaganda detection. Their schema introduced 18 propaganda techniques (e.g. loaded language, name-calling, appeal to fear) annotated at the span level, which became standard ever since, with minimal revisions (e.g. condensing the schema to 10 (Kyslyi et al., 2025) or 14 labels by Da San Martino et al. (2020); Sprenkamp et al. (2023); Abdullah et al. (2022)). At the same time, subsequent work indicated high subjectivity and complexity of the annotation process (Maarouf et al., 2024; Sahitaj et al., 2025).
Propaganda classification. Recent work has explored LLMs for propaganda detection. Sahitaj et al. (2025) proposed a hybrid annotation framework that leverages LLMs to pre-annotate texts, followed by human validation. This approach uses a hierarchical taxonomy of 14 techniques grouped into three broader categories, and demonstrates that LLM pre-annotations improve both inter-annotator agreement and annotation efficiency. Jose and Greenstadt (2025) evaluated GPT-3.5, GPT-4, and Claude on six techniques in news articles, while Hasanain et al. (2023) used GPT-4 to generate multilabel and sequence tagging annotations for 23 techniques in Arabic. Sprenkamp et al. (2023) benchmarked GPT-3 and GPT-4 for multi-label classification, showing close to SOTA performance. Recent work has begun to move beyond surface-level detection toward deeper analysis of propaganda (Liu et al., 2025) and LLM explanations (Hasanain et al., 2025). In contrast to these approaches, our work focuses on how annotation schema design and prompting strategy interact to shape learnability and robustness in high-level propaganda classification.

3 Methodology and Experimental setup

3.1 Datasets and Annotations

We used the HQP benchmark dataset (Maarouf et al., 2024) to co-annotate tweets from Sahitaj et al. (2025). It contains 30,000 English tweets annotated

[1] https://github.com/verosol/propaganda_hierarchical
[2] https://huggingface.co/collections/xplainlp/propaganda-classification

with fine-grained propaganda techniques, collected at the start of the Russian war on Ukraine.

To capture the dynamics of modern online propaganda, we developed a new set of labels using an iterative, mixed-methods approach combining expert annotation, LLM-assisted harmonization, and group consensus. Five annotators with advanced backgrounds in computer science and familiarity with propaganda research, and independently labeled 300 tweets from the HQP dataset, describing the intent of each message in free text rather than selecting from predefined categories.

These descriptions were clustered using GPT-4 to identify overlapping concepts and refined in group discussion, where we additionally grounded the resulted categories in Starbird (2019) and Entman (1993) works. This resulted in 17 fine-grained labels, which were also organized into five higher-level categories, forming a two-tier taxonomy (Table 9 and 8; Appendix D for the complete label set). We then validated and refined the schema on a larger sample of 500 tweets, where each tweet was annotated on three levels: (i) all fine-grained techniques present, (ii) the single most prominent technique, and (iii) the overarching high-level intent. This process yielded the final hierarchical labeling schema, designed to support both nuanced analysis and robust high-level classification.

3.2 Inter-Annotator Agreement

To gauge annotation reliability, we computed inter-annotator agreement (IAA) on the propaganda portions of the golden test sets. As shown in Table 1, our schema yielded only moderate agreement (κ=0.309, α=0.308), whereas Sahitaj et al. (2025)'s schema achieved substantially higher agreement (κ=0.594, α=0.594).

The gap highlights a fundamental trade-off. Sahitaj et al. (2025)'s categories are easier to apply consistently, yielding stronger agreement. By contrast, our intent-focused taxonomy is cognitively more demanding and introduces genuine ambiguity: annotators often disagreed not because of unclear guidelines, but because multiple propagandistic intents were simultaneously plausible. This reflects the inherent difficulty of labelling intent in short, noisy social media texts. Across framing-style annotation campaigns, to which our scheme is arguably closer than to traditional propaganda taxonomies, reported raw inter-annotator agreement is typically low to moderate (e.g., Krippendorff's $\alpha \approx 0.3$–0.4 for generic frames and ≈ 0.1–0.2 for span-level unitizing; Card et al., 2015; Piskorski et al., 2023; Bassi et al., 2025). Prior work consistently attributes this to the inherent subjectivity and multi-label nature of framing, and treats IAA primarily as a signal for managing annotators and consolidation rather than a strict quality threshold.

Schema	**Cohen's κ**	**Krippendorff's α**
Our scheme	0.309	0.308
Sahitaj et al.	0.594	0.594

Table 1: Overall inter-annotator agreement on propaganda tweets (main labels).

Metric	Score
Overall overlap (A1–A4)	0.66
Krippendorff α_{nom}	0.78
A5 vs. majority vote macro-F1	0.93
A5 vs. majority vote micro-F1	0.91

Table 2: Validation of clustered high-level schema on 100 tweets. A1–A4: human annotators; A5: clustering-based labels. Strict uses the primary label only; lenient counts agreement if either primary or secondary label matches.

3.3 Schema Comparison

We consider two schemas, each with **17** propaganda techniques plus one non-propaganda class, grouped into **six high-level** categories.

Sahitaj et al. (2025)'s schema is a span-based taxonomy developed for news and adapted to Twitter. For comparability at the high level, we *derive* six umbrellas by clustering their 17 propaganda techniques + one non-propaganda technique using a similarity matrix combining label co-occurrence (Jaccard) and Dawid–Skene (Dawid and Skene, 1979) posterior profiles; we fix the non-propaganda label to its own group (details in App. B). Four annotators (A1-A4) then re-labelled a 200-tweet gold set with these six categories while seeing the original span-level annotations. We obtain substantial strict agreement (overall overlap 0.66, Krippendorff's $\alpha_{\text{nom}} = 0.78$). The clustering-based labels (A5) match the human majority vote with macro-/micro-F1 of $0.93/0.91$ (Tab. 2), indicating that the induced high-level schema is stable and reproducible.

3.4 Models

All experiments use **(i) GPT-4.1-nano**(OpenAI, 2023) (release: 2025-04-14) base and fine-tuned with the OpenAI API and provider-standard optimization defaults; and **(ii) Phi-4 14B**(Abdin et al., 2024) models, base and fine-tuned with LoRA-adapters; (iii-iv) Qwen2.5-14B (Yang et al., 2024) and Qwen3-14B (Yang et al., 2025) base and fine-tuned. The fine-tuning details are documented in Appendix C. For each schema, we used a fixed split totaling **500 tweets** (all drawn from the HQP corpus): training set (210 tweets, with 105 propaganda, 105 non-propaganda); validation set (90 tweets, with 45 propaganda, 45 non-propaganda); golden test set (200 tweets, with 100 propaganda, 100 non-propaganda).

Non-propaganda items are identical across schemas, while propaganda items differ to preserve balanced label distribution, but all tweets are drawn from the same source corpus. The golden test set is balanced at the binary level, but distributions across fine-grained propaganda techniques remain slightly skewed.

3.5 Prompting strategies

We compare: (i) **Direct-High** (predict a high-level category directly from text + category descriptions) and (ii) our **HiPP** method, which is **Main→High** (first predict the prominent technique, then the high-level category in the same pass, conditioned on the main prediction).

Evaluation. We report macro-F1 and weighted-F1. Macro-F1 highlights performance on minority classes, which is critical under the imbalanced technique distributions observed in HQP; weighted-F1 reflects prevalence.

4 Results

We evaluate models across both schemas and both prompting strategies. Table 3 reports results for both baseline (zero-shot) and fine-tuned conditions. Across the board, fine-tuning substantially improves absolute performance, while preserving the qualitative trends observed in the baselines. Three main patterns emerge.

Effect of fine-tuning and prompting. Fine-tuning is decisive for all configurations: it yields gains of $+0.09$–$+0.30$ weighted-F1 compared to zero-shot. **HiPP** (Main→High) tends to underperform Direct-High in the Base regime, especially on our lower-IAA schema, but benefits disproportionately from fine-tuning. For the Sahitaj schema, FT Main→High is the best configuration for *every* backbone. On our schema, FT Main→High is clearly best for GPT-4.1-nano and competitive for the other models, while the stronger configurations obtain their top scores with FT Direct-High. These results indicate that the effectiveness of HiPP depends strongly on the interaction between supervision regime and label-space characteristics. In the zero-shot (Base) setting, Main→High often underperforms Direct-High, suggesting that hierarchical decomposition can amplify noise when model representations are not adapted. After fine-tuning, however, Main→High consistently improves or matches Direct-High, particularly on the lower-IAA schema, indicating that hierarchical prompting becomes beneficial once the model has learned task-specific representations. Rather than depending solely on model capacity, the observed gains arise from the interaction between supervision and label reliability. Overall, these results show that hierarchical prompting is not universally beneficial, but relies on fine-tuning to provide gains.

Effect of schema / IAA. Absolute performance is consistently higher on the Sahitaj schema across all models and strategies, with gaps of roughly $+0.05$–$+0.15$ in weighted-F1. This aligns with its substantially stronger inter-annotator agreement and supports our claim that our intent-based taxonomy is the more challenging setting. The fact that Main→High excels on both schemata after fine-tuning indicates that the **HiPP** strategy provides a robust methodological advantage and is not merely overfitting to a specific label design. Even under strong annotator disagreement, the fine-tuned Main→High model can recover the majority-vote label. For one tweet, annotators split between Shift Blame (2 votes), Distort Reality (2), and Delegitimise (3), yielding Krippendorff's $\alpha = 0.18$. Only the Main→High configuration predicted the majority label correctly; the full example is given in Appendix A.

Model comparison. Among the open-weight models, the two Qwen models achieve the highest scores overall, with Qwen3-14B reaching 0.685 W-F1 on our schema and 0.661 on Sahitaj. Phi-4 14B consistently outperforms GPT-4.1-nano in both regimes and on both schemata, confirming that our findings are not specific to a single model family but hold across multiple moderately sized

Schema	Regime	Strategy	GPT-4.1-nano		Phi-14B		Qwen2.5-14B		Qwen3-14B	
			M	W	M	W	M	W	M	W
Ours	Base	Direct	0.204	0.379	0.379	0.554	0.301	0.536	0.334	0.483
	Base	M→H	0.137	0.225	0.281	0.399	0.259	0.457	0.189	0.333
	FT	Direct	0.233	0.477	0.375	**0.562**	**0.431**	**0.630**	**0.539**	**0.685**
	FT	M→H	**0.376**	**0.571**	**0.412**	0.554	0.416	0.605	0.431	0.614
Sahitaj	Base	Direct	0.271	0.405	0.421	0.439	0.479	0.554	0.424	0.519
	Base	M→H	0.284	0.429	0.364	0.195	0.360	0.192	0.374	0.373
	FT	Direct	0.441	0.574	0.493	0.511	0.498	0.612	0.448	0.597
	FT	M→H	**0.529**	**0.623**	**0.548**	**0.602**	**0.560**	**0.657**	**0.560**	**0.661**

Table 3: Results across configurations. Scores are macro-F1 (M) and weighted-F1 (W). Strategies: Direct = Direct-High, M→H = Main→High. Regimes: Base = zero-shot baseline, FT = fine-tuned. For Sahitaj et al. (2025) the **HiPP** Main→High strategy consistently gives the best results after fine-tuning, and absolute scores are higher overall due to stronger IAA.

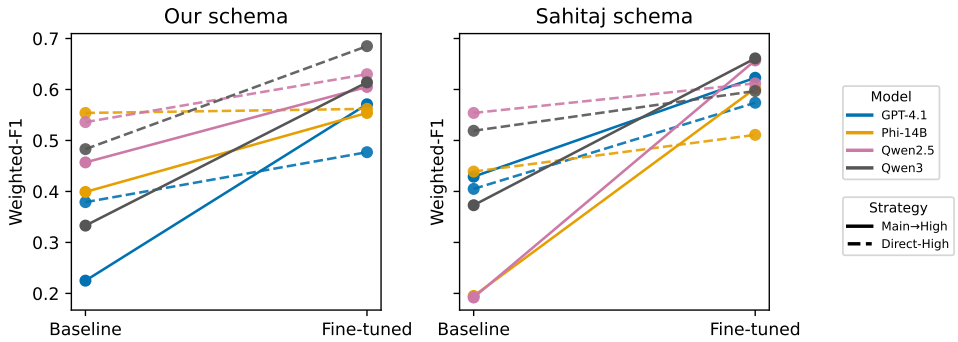


Figure 1: **Weighted-F1 improvements from Direct-High to Main-High.** For each backbone, lines connect zero-shot (Baseline) and fine-tuned scores. On both schemata, Main→High lags behind Direct-High in zero-shot but gains more from fine-tuning, especially on our lower-IAA taxonomy. (GPT-4.1 here refers to the GPT-4.1-nano variant.)

LMs. Weighted-F1 scores are systematically higher than macro-F1, reflecting the skewed label distribution: frequent classes dominate weighted-F1, while macro-F1 highlights the remaining difficulty in rarer intents.

5 Error Analysis

5.1 Setup

We perform a qualitative error analysis that combines confusion matrices with lemma-frequency statistics. For the latter, we lemmatise all test tweets using the spaCy large English [3], then group them into true positives, true negatives, false positives, and false negatives, separately for each model, training regime (Base vs. FT), and schema (ours vs. Sahitaj et al. (2025)). For every confusion-cell we compute the most frequent lemmatised tokens. Full confusion matrices and lemma tables are provided in Github repository and example for Qwen3 results is given in Appendices E and F.

[3] `en_core_web_lg` (`https://spacy.io/models/en`)

We also analyse our best results for Sahitaj and our scheme (both Qwen3-14B), by visualising all confusion cells (true, predicted) as TF–IDF vectors and projecting them to two dimensions with PCA (Figs. 2; we include all cells, even those with only a single instance). Since we are interested in the global variance within the datapoints, we apply PCA instead of other techniques like t-SNA or UMAP.

Our analysis focuses on two research questions: (i) non-propaganda vs. propaganda confusions, which dominate the error mass, and (ii) within-propaganda confusions between high-level intent and technique categories.

5.2 Non-Propaganda vs. Propaganda

Across all models, schemas and regimes, correctly and incorrectly classified non-propaganda tweets share the same topical lexicon: ukraine, russia, war, people, crisis are among the most frequent lemmas in true non-propaganda (0→0). At the same time, almost identical vocabularies appear

in non-propaganda tweets that are misclassified as propaganda. The key difference is strength of framing. Mislabelled instances contain additional high-intensity cues such as nazi/nazis, genocide, terrorist, destroy, ban, closethesky, sendnatotoukraine, or insults and culture-war terms (e.g., fuck, woke).

In the Base Direct-High regime, this leads all four models to be hyper-sensitive to these keywords: any combination of ukraine/russia/war with Nazi/WWII rhetoric is frequently mapped to attack/shift-blame/emotional category, even when the tweet is reporting or critical commentary. In the Base Main→High setting, this effect is amplified: once a main technique is chosen, war+Nazi lexicon tends to be "locked in" as a propaganda high-level label, inflating 0→propaganda errors.

Fine-tuning, and especially fine-tuned Main→High, sharply reduces this behaviour. For both schemas, the first row of the confusion matrices becomes much more diagonal: neutral war reporting, often including URLs and event descriptors, is kept in class 0, while the remaining 0→propaganda mass concentrates on borderline, highly polarised cases such as explicit genocide claims or accusations of neo-Nazi identity. This pattern holds for GPT-4.1-nano, Phi-4, Qwen2.5 and Qwen3.

5.3 Confusion between propaganda classes

5.3.1 Our Schema

For our intent and framing based schema, propaganda labels (1–5) are mainly confused with adjacent goals in the hierarchy. Distort-Reality tweets misclassified as Shift Blame (1→2) typically mix historical reinterpretation (coup, election, proxy war, Donbass) with responsibility-shifting (NATO, Biden, West), making the boundary between revising the past and justifying current aggression inherently fuzzy. Shift-Blame tweets mislabelled as Delegitimise (2→3) are dominated by dehumanising rhetoric (nazi, battalion, hitler, swastika, salute), so models focus on moral condemnation rather than the blame-transfer structure. Manufacture-Consent posts mispredicted as Shift or Delegitimise (4→2/3) combine in-group mobilisation (istandwithputin, istandwithrussia, bandera, ukronazis) with explicit attacks on the out-group. Models sometimes prioritise the attack aspect over the identity-building goal. Confuse & Distract (5) exhibits the most heterogeneous lexicon (paypal, ban, twitter, donation, tank, weapon, western media), and its instances leak into several other propaganda categories. This suggests that cues for distraction and fragmentation (multiple topics, logistics, meta-commentary) are not well captured by simple lexical patterns.

Fine-tuning makes these confusions more taxonomically local. In FT Direct-High, correctly classified Shift-Blame items concentrate on expected frames (nazi, nato, victimhood), while misclassified ones tend to mix elections, coups and moral condemnation. In FT Main→High, off-diagonal mass is largely restricted to Distort↔Shift and Shift↔Delegitimise; long-range errors (e.g., Distort↔Confuse) are strongly reduced.

5.3.2 Sahitaj scheme

For the clustered Sahitaj schema, the high-level groups already merge multiple techniques, and lemma-based errors reflect multiple tendencies. Non-propaganda is often mislabelled as Emotional & Loaded Persuasion or Deflections & Distractions. This is driven lexicon such a *nazi/nazis, threat, destroy, bomb, ban, closethesky* and other standard war lexicon. Within propaganda classes, confusions concentrate on (i) Emotional ↔ Deflections, when strong affect co-occurs with causal or whataboutist reasoning (*nazi, propaganda, fake, fear, attention*) and (ii) Deflections ↔ Argument Manipulations, when reframing causality is combined with thought-terminating clichés or strawman attacks.

Compared to our schema, the clustered Sahitaj categories show cleaner lexical separations. Popularity and patriotic appeals are tightly tied to slogan hashtags (istandwithrussia, standwithputin, sendnatotoukraine), while Argument Manipulations collect more abstract reasoning tokens (logic, lie, believe, question). After fine-tuning, Qwen2.5 and Qwen3 in particular handle the patriotic class almost perfectly, with very few leaks into other clusters.

5.4 Model and Regime Effects

Phi-4's errors are lexically the sharpest. When it misclassifies non-propaganda as Deflections, the top tokens overwhelmingly support that reading (threat, propaganda, nato, stop, biden). GPT-4.1-nano is the noisiest, spreading the same war vocabulary across more categories. Qwen2.5 and Qwen3 sit in between: their error lexicons are more concentrated than GPT-4.1-nano's, but still less focused than Phi-4's. Qwen3 is slightly more conserva-

tive post-FT, with especially clean non-propaganda rows and crisp handling of slogan-driven patriotic appeals. Base models are much more sensitive to simple lexical cues such as “war" and "Nazi”, regardless of context. Fine-tuning shifts errors toward genuinely ambiguous, polarised tweets and shrinks lemma distributions in error cells: generic war terms recede, while extreme cues (*nazi, genocide, ban, closethesky*) dominate the remaining mistakes.

In zero-shot, Main → High increases false positives. Once a main technique is predicted, the model tends to “commit” to a propaganda label whenever strong lexical cues are present. After fine-tuning, the same hierarchical structure becomes stabilising, since the results show that non-propaganda is rarely assigned any technique at the MAIN level, and HIGH-level errors align with conceptual neighbourhoods (Distort-Shift, Emotional ↔ Deflections ↔ Argument).

5.5 Performance of clustering-based labels

The clustered high-level labels derived from Sahitaj et al. (2025) turn out to be both annotation- and model-friendly. On the human side, they achieve substantial strict agreement and very high lenient agreement in our human validation, indicating that annotators can reliably reason when given the underlying span-level techniques. On the modelling side, several clusters behave strikingly well across all models and regimes. Patriotic & Catchy Appeals is almost a “slogan class”: after fine-tuning, tweets with dense campaign-style hashtags (e.g., *istandwithputin*, *istandwithrussia*) are classified with very high precision and recall. No Propaganda is the main beneficiary of fine-tuning with **HiPP**: while zero-shot models often over-label war content as propagandistic, FT Main→High yields a strong diagonal for this class, keeping neutral, URL-rich reporting in NP and relegating only the most polarised war+Nazi posts to propaganda. Within propaganda, Emotional & Loaded Persuasion, Deflections & Distractions, and Argument Manipulations form a tight triad: they achieve good F1 scores, and their remaining errors are largely confined to confusions inside this trio (e.g., descriptive Emotional content drifting into Deflections, or reasoning-heavy Deflections drifting into Argument Manipulations). Taken together with the high co-annotator agreement, this suggests that the clustered Sahitaj schema is coarse enough to be robust and learnable, yet structured enough to preserve meaningful distinctions in propagandistic style.

5.6 PCA analysis

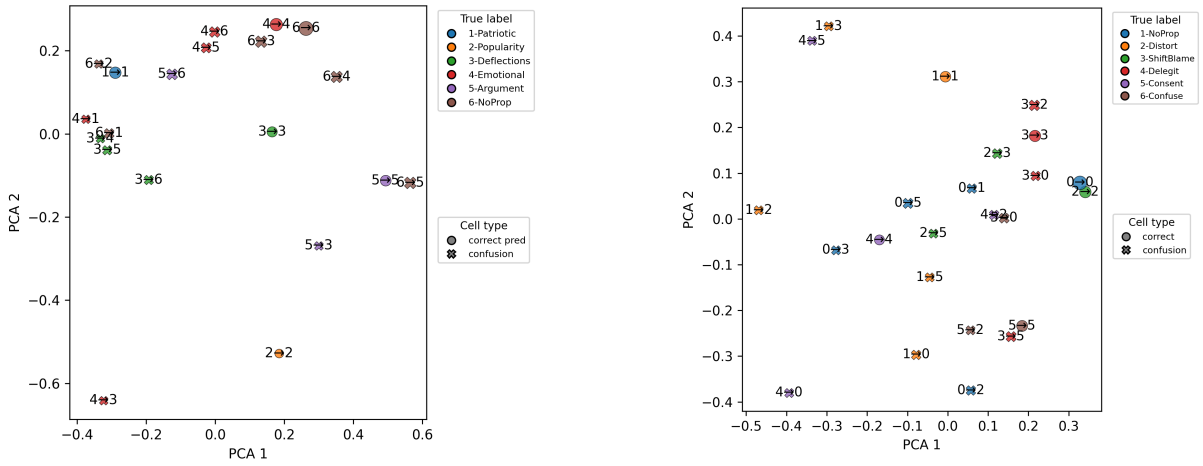
In both schemata, cells sharing the same true label form compact lexical neighbourhoods, and confusions stay close to the corresponding correct cell, indicating that errors are locally coherent rather than arbitrary. For the clustered Sahitaj labels, these neighbourhoods are tight and well separated, whereas for our schema the Distort, ShiftBlame, Delegitimise and Consent intents occupy a broad overlapping region with many small cross-label cells. In plot B, we can observe that NoProp (0) cells lie adjacent to correctly predicted ShiftBlame and Delegitimise cells (2→2, 3→3), and its false positives (0→1/2/3/5) occupy the same region, illustrating how borderline non-propaganda tweets share vocabulary with attack and responsibility-shifting propaganda. The propaganda intents Distort, ShiftBlame, Delegitimise and Consent (1–4) form a broad overlapping cluster. Confuse & Distract (5) has no clean cluster of its own, as its cells are interspersed with ShiftBlame and Delegitimise, reflecting the heterogeneous, catch-all nature of this intent. This denser view confirms our qualitative finding that the Sahitaj clusters are easier and lexically cleaner, while our taxonomy induces a more tangled, higher-variance decision space.

6 Discussion

Our study examined three dimensions of propaganda detection: model portfolio selection, schema-level comparison, and prompting strategy. Across all experiments, a consistent pattern emerges: fine-tuning is decisive, and the **HiPP** Main→High strategy is particularly beneficial once models have been adapted to the task.

Zero-shot performance varies widely across models and strategies, whereas fine-tuning reliably turns weak baselines into competitive systems and makes methodological effects visible that are almost invisible in the Base regime. This suggests that base models substantially underestimate the potential of open-weight LMs for specialised tasks such as propaganda detection.

Schema design and reliability also shape outcomes. Models trained on Sahitaj et al. (2025)’s higher-agreement schema (κ=0.594) consistently outperform those trained on our intent-focused, lower-agreement taxonomy (κ=0.309). This illustrates a fundamental trade-off: simplified,



(a) Sahitaj high-level clusters

(b) Our intent-based high-level schema

Figure 2: PCA of confusion-cell TF–IDF vectors for Qwen3-14B in its best setting. Each point represents a (true, predicted) high-level label cell; colours indicate true labels and markers distinguish correct vs. confused cells.

technique-centric schemata provide clearer learning signals, while more ecologically realistic, intent-based schemata better reflect real-world complexity but introduce ambiguity and label noise. However, the best settings only differ in 0.02 points making both schema application possible, depending on the application use case, without strong performance trade-off.

Among the open-weight models we test, Phi-4 14B already outperforms GPT-4.1-nano across schemata, and the two Qwen models achieve the best absolute scores, with Qwen3-14B reaching 0.685 weighted-F1 on our schema and 0.661 on Sahitaj. The benefits of HiPP are most pronounced after fine-tuning and in the presence of noisier label spaces, rather than being determined solely by backbone strength: for GPT-4.1-nano, fine-tuned Main→High outperforms Direct-High on both schemata, improving weighted-F1 by +0.094 (ours) and +0.049 (Sahitaj). For stronger models, Main→High remains the best configuration on the Sahitaj schema and is competitive on ours, while some models obtain slightly higher scores with Direct-High. Overall, this pattern suggests that **HiPP** provides a useful inductive bias for coping with noisy, imbalanced annotations, especially when the label space is challenging or model capacity is limited, without requiring changes to the underlying schema. However, as both Direct-High and Main→High observe the same supervision signal, performance gains can be attributed to hierarchical decomposition rather than increased label information.

Overall, the analysis suggests that all four models, across both taxonomies, implement a similar inductive bias: they strongly over-weight topic + affective markers (Ukraine/Russia/war + Nazi/WWII rhetoric) as signals of propaganda, and they treat high-level propaganda intents in a locally coherent way, mostly confusing neighbouring goals rather than arbitrarily jumping across the taxonomy. Fine-tuning in the **HiPP** Main→High regime turns these models from crude lexical detectors into structured classifiers whose residual mistakes are concentrated in the most contentious cases—precisely where human annotators also tend to disagree.

At the same time, our intent-based schema remains visibly harder than the clustered Sahitaj schema. Even after fine-tuning, several of our propaganda classes remain mutually confused, mirroring the lower human IAA we observe. From a practical perspective, this highlights a trade-off between expressive taxonomies and clustered high-level schema, which is easier to learn and deploy but less informative, making it less suitable for application on real-world data. Our results indicate that, when fine-tuned with HiPP, modern LLM models from different families converge to very similar and interpretable error structures, providing some reassurance that conclusions drawn from one model reasonably transfer to others.

The PCA visualisations of confusion cells from the best settings also highlight how our taxonomy reshapes the decision space. While the Sahitaj clusters yield compact, well-separated lexical neighbourhoods with mostly local errors, our intent-

based schema produces a much denser, overlapping landscape, suggesting that the added expressivity of our labels comes at the cost of a more difficult and inherently noisier classification problem. In this context, we believe that our proposed taxonomy has substantial value. Beyond methodological findings, our intent-focused taxonomy opens new opportunities for research and practice. By emphasizing communicative goals rather than surface-level techniques, it enables nuanced studies of how propaganda strategies evolve across conflicts and cultures, and how they interact with public opinion and polarization. For practitioners, it can support early-warning systems that focus not just on whether a message is propaganda, but also why it was crafted and how it seeks to influence audiences. At the same time, high level labels we created for Sahitaj et al. scheme, show the most promise in terms of IAA. Although validated on a limited sample, they show one of the highest agreement results seen so far in literature.

7 Conclusion

In this work, we introduced a new hierarchical labeling schema for online propaganda, focused on intent and key message, and conducted a controlled comparison between our taxonomy and the recent framework by Sahitaj et al. (2025). We evaluated how schema design, annotation reliability, and prompting strategies interact, comparing various similar sized models, direct high-level classification and a **HiPP** approach. Our results show that while schema reliability shapes performance, LLM fine-tuning and **HiPP** provides an advantage, offering robust supervision specifically when annotation quality is imperfect or datasets are heterogeneous.

Limitations

Our study has several limitations. First, it was conducted on relatively small subsets of the HQP corpus (500 items per schema), with results based on a single split. Testing across two schemas and two models helped mitigate setup-specific variance, but future work should confirm stability under multiple random splits. Given the relatively small dataset size and single train/test split, the reported results should be interpreted as indicative of methodological trends rather than definitive performance estimates. Second, our clustering of Sahitaj et al.'s (Sahitaj et al., 2025) main labels into high-level groups involved interpretive choices (e.g., λ for similarity weighting). Future work should therefore scale up evaluation, test cross-validation setups, and replicate high-level clustering with alternative methods or expert-driven designs. It will also be valuable to explore hierarchical prompting in multi-task or multi-step architectures, and to design hybrid corpora that balance ecological realism with annotation reliability.

Exploring additional open-weight and proprietary models and their larger and smaller versions would help better generalize the trends we found in this study.

Finally, our labeling schema operates at the intersection of *framing* and *propaganda*. Future work will focus on disentangling these two phenomena within one schema to better estimate which part is more challenging for LLMs.

Our annotators were primarily early-career researchers with backgrounds in computer science from Germany and Ukraine; while this provided valuable regional expertise, broader demographic coverage would further strengthen future work.

Ethical Considerations

The annotation process inherently reflects subjective judgments, and schema design choices may encode cultural or political biases. To mitigate this, we used a multilingual team of seven annotators and conducted consensus discussions. Nevertheless, our models may still exhibit bias toward particular geopolitical or linguistic contexts, and results should be interpreted with caution.

Finally, while our methods are intended to support fact-checking and disinformation monitoring, they could be misused for large-scale surveillance or censorship. We emphasize that detection tools should always be deployed in combination with human oversight and transparent governance structures.

While working on the manuscript, we have used LLM assistants for the purpose of spell-checking and as a writing assistant, but not for the creation of the content.

Acknowledgments

The work on this paper was performed in the scope of the projects "VeraXtract" (16IS24066) and "news-polygraph" (reference: 03RU2U151C) funded by the German Federal Ministry for Research, Technology and Aeronautics (BMFTR).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Malak Abdullah, Ola Altiti, and Rasha Obiedat. 2022. Detecting propaganda techniques in english news articles using pre-trained transformers. In *2022 13th International Conference on Information and Communication Systems (ICICS)*, pages 301–308. IEEE.

Davide Bassi, Dimitar Iliyanov Dimitrov, Bernardo D'Auria, Firoj Alam, Maram Hasanain, Christian Moro, Luisa Orrù, Gian Piero Turchi, Preslav Nakov, and Giovanni Da San Martino. 2025. Annotating the annotators: Analysis, insights and modelling from an annotation campaign on persuasion techniques detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17918–17929, Vienna, Austria. Association for Computational Linguistics.

Judit Bayer, Bernd Holznagel, Katarzyna Lubianiec, Adela Pintea, Josephine B Schmitt, Judit Szakács, and Erik Uszkiewicz. 2021. Disinformation and propaganda: impact on the functioning of the rule of law and democratic processes in the eu and its member states. *European Union*.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association 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, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024. Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2724–2744, Torino, Italia. ELRA and ICCL.

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

Maram Hasanain, Md Arid 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.

Julia Jose and Rachel Greenstadt. 2025. Are large language models good at detecting propaganda? *arXiv preprint arXiv:2505.13706*.

Garth S Jowett and Victoria O'donnell. 2018. *Propaganda & persuasion*. Sage publications.

Roman Kyslyi, Nataliia Romanyshyn, and Volodymyr Sydorskyi. 2025. The unlp 2025 shared task on detecting social media manipulation. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 105–111, Online. Association for Computational Linguistics.

Jiateng Liu, Lin Ai, Zizhou Liu, Payam Karisani, Zheng Hui, Yi Fung, Preslav Nakov, Julia Hirschberg, and Heng Ji. 2025. PropaInsight: Toward deeper understanding of propaganda in terms of techniques, appeals, and intent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5607–5628, Abu Dhabi, UAE. Association for Computational Linguistics.

Abdurahman Maarouf, Dominik Bär, Dominique Geissler, and Stefan Feuerriegel. 2024. Hqp: A human-annotated dataset for detecting online propaganda. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6064–6089. Association for Computational Linguistics.

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*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

E Perez. 2022. Strategic disinformation: Russia, ukraine and crisis communication in digital era.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Ariana Sahitaj, Premtim Sahitaj, Veronika Solopova, Jiaao Li, Sebastian Möller, and Vera Schmitt. 2025. Hybrid annotation for propaganda detection: Integrating LLM preannotations with human intelligence. In *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pages 215–228, Vienna, Austria. Association for Computational Linguistics.

Veronika Solopova, Christoph Benzmüller, and Tim Landgraf. 2023. The evolution of pro-kremlin propaganda from a machine learning and linguistics perspective. In *Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP)*, pages 40–48, Dubrovnik, Croatia. Association for Computational Linguistics.

Veronika Solopova, Viktoriia Herman, Christoph Benzmüller, and Tim Landgraf. 2024. Check news in one click: NLP-empowered pro-kremlin propaganda detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 44–51, St. Julians, Malta. Association for Computational Linguistics.

Veronika Solopova, Robert Nickel, and Dorothea Kolossa. 2025. *Lagging Behind: Challenges of Adapting Automated Propaganda Detection to the Chronological and Geographic Evolution of Language*, chapter Chapter 9.

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

Ivan Srba, Olesya Razuvayevskaya, João A Leite, Robert Moro, Ipek Baris Schlicht, Sara Tonelli, Francisco Moreno García, Santiago Barrio Lottmann, Denis Teyssou, Valentin Porcellini, and 1 others. 2024. A survey on automatic credibility assessment of textual credibility signals in the era of large language models. *arXiv preprint arXiv:2410.21360*.

Kate Starbird. 2019. Disinformation's spread: Bots, trolls and all of us. *Nature*, 571(7766):449.

Prashanth Vijayaraghavan and Soroush Vosoughi. 2022. Tweetspin: Fine-grained propaganda detection in social media using multi-view representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3433–3448.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

An Yang, Yuxiao Yang, Qingyao Ai, and 1 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115. Describes the Qwen2.5 model family, including Qwen2.5-14B.

An Yang, Yuxiao Yang, Qingyao Ai, and 1 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388. Describes the Qwen3 model family, including Qwen3-14B.

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

A Representative Low-IAA Example

Tweet (full).

> *@USER @USER Oh no ....another "Jews for Nazis". Is it because that the Nazis are killing Russians and not Jews — makes it OK?*

Human votes (7 raters). 2× *Distort reality and rewrite the past* (1), 2× *Shift blame and Justify Aggression* (2), 3× *Delegitimize the opponent* (3). The plurality gold label is therefore *Delegitimize the opponent* (3/7).

Model predictions. Table 4 compares system outputs against the gold label. Only fine-tuned Main→High predicts the majority-vote label, while all other strategies misclassify the tweet.

Model	Pred (int)	Predicted label
Direct (Base)	2	Shift blame & Justify Aggression
Direct (FT)	2	Shift blame & Justify Aggression
M→H (Base)	1	Distort reality & Rewrite past
M→H (FT)	**3**	**Delegitimize opponent**
Gold (plurality)	**3**	**Delegitimize opponent**

Table 4: **Representative low-IAA case.** Annotator votes: 1. Distort reality (2/7 votes), 2. Shift blame (2/7 votes), 3. Delegitimize opponent (3/7 votes). Plurality gold = *Delegitimize opponent* (3/7). Only fine-tuned M→H predicts the correct label.

B Clustering of Sahitaj et al.'s Techniques

To enable high-level comparisons, we clustered the 17 main propaganda techniques from Sahitaj et

al. (Sahitaj et al., 2025) into five umbrella propaganda categories plus the non-propaganda class.

Method. We construct a similarity matrix $S_{ij} = \lambda J_{ij} + (1 - \lambda) DS_{ij}$, where J_{ij} is Jaccard co-occurrence across tweets and DS_{ij} is the similarity of Dawid–Skene posteriors. The non-propaganda label is fixed to its own group. To avoid data leakage, clustering was performed only on the training set.

Agreement Across λ. Table 5 shows agreement scores across similarity weightings. Agreement improved as λ approached pure co-occurrence ($\lambda = 1.0$), peaking at Fleiss' $\kappa = 0.738$ and Krippendorff's $\alpha = 0.738$. We selected $\lambda = 0.5$ as a balanced setting: clusters remained interpretable and well-sized, with only a small drop in agreement.

λ	Fleiss' κ	Krippendorff's α
0.0	0.683	0.684
0.1	0.714	0.714
0.2	0.719	0.719
0.3	0.719	0.719
0.4	0.719	0.719
0.5	0.719	0.719
0.6	0.719	0.719
0.7	0.725	0.725
0.8	0.725	0.725
0.9	0.733	0.733
1.0	0.738	0.738

Table 5: Agreement vs. λ for hybrid similarity weighting on Sahitaj et al.'s schema.

Cluster Compositions. The $\lambda = 0.5$ solution yielded six interpretable umbrella categories. These roughly align with communicative functions such as patriotic appeals, popularity appeals, deflections, emotional persuasion, logical/argument manipulations, and non-propaganda.

- Group 1: {15, 5}
- Group 2: {16, 9}
- Group 3: {14, 17, 7, 8}
- Group 4: {12, 13, 3}
- Group 5: {1, 10, 11, 2, 4, 6}
- Group 6: {18 (non-propaganda)}

C Fine-tuning Details

We document our fine-tuning strategies in Table 6 for the GPT-4.1-nano model (fine-tuned via OpenAI API) and in Table 7 for the Phi4-14B model (fine-tuned via the unsloth framework [4]). All results are based on a single fixed split (no cross-validation). Fine-tuning ran to completion without early stopping.

Fine-tuning config	
Model (base)	GPT-4.1-nano (OpenAI) (release: 2025-04-14)
epochs	3
batch_size	1
learning rate multiplier	0.1

Table 6: Details for fine-tuning GPT-4.1-nano

Fine-tuning config	
batch_size	4
gradient_accumulation_steps	4
warmup_steps	5
num_train_epochs	5
learning_rate	2e-4
optim	"adamw_8bit"
weight_decay	0.01
lr_scheduler_type	"linear"
seed	3407
LoRA / PEFT Adapter	
r	16
target_modules	{q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj}
lora_alpha	16
lora_dropout	0
bias	"none"
use_gradient_checkpointing	"unsloth"
random_state	3407
use_rslora	False
loftq_config	None
Model load / unsloth configuration	
from_pretrained(model_name)	"unsloth/Phi-4"
full_finetuning	False
load_in_4bit	False
Final Quantization	8 bit

Table 7: Hyperparameters for fine-tuning Phi4-14B, including LoRA adapter configuration

We fine-tuned Phi-4 on 1 GPU Nvidia A6000 48Gb for approx. 2 hours each experiment, while gpt-4.1nano was fine-tuned on the OpenAI servers. The whole set of gpt-4.1nano experiments cost 40 US dollars.

[4] https://github.com/unslothai/unsloth

D New Labelling schema

For completeness, we provide the full set of labels defined in our intent-focused hierarchical taxonomy. The schema consists of 17 fine-grained propaganda techniques plus one non-propaganda class, which are grouped into five high-level propaganda categories and one non-propaganda category. Table 9 and Tables 8 restates the label inventory introduced in the main text.

Compared to prior taxonomies such as Da San Martino et al. (2019, 2020) and the hybrid hierarchical schema of Sahitaj et al. (2025), our intent-based framework shifts focus from surface-level rhetorical techniques to the underlying communicative goals of propaganda. Earlier taxonomies primarily captured stylistic and linguistic devices (e.g., loaded language, name-calling, appeal to fear), which describe how persuasion is achieved but not why a message was crafted. By contrast, our schema introduces high-level intent categories such as Shift Blame and Justify Aggression, Manufacture Consent and Identity, and Confuse and Distract, which unify multiple lower-level techniques under shared strategic purposes. This design captures manipulative intents, such as reframing moral responsibility, generating uncertainty, or mobilising identity-based polarisation, that older schemas either scattered across labels or missed entirely. As a result, the new taxonomy provides a more explanatory and goal-oriented lens for studying online propaganda, facilitating cross-context comparisons and supporting interpretability in LLM-based detection systems.

High-Level Categories
1. Distort Reality and Rewrite the Past Goal: Undermine truth and legitimize the present through distortion. Subtypes: 2, 11, 15, 17 (False Origin Attribution, Circular Reasoning, Appeal to Ignorance). Rationale: Denial, fabrication, and invented evidence challenge fact-based narratives.
2. Shift Blame and Justify Aggression Goal: Reframe aggressor as victim or rational actor. Subtypes: 4, 5, 13, 17 (False Causality, False Balance). Rationale: Inverts moral frameworks to rationalize wrongdoing or redirect guilt.
3. Delegitimize the Opponent Goal: Undermine credibility and morality of adversaries. Subtypes: 1, 3, 8, 9, 12, 6. Rationale: Frames enemies as liars, extremists, or inhuman entities.
4. Manufacture Consent and Identity Goal: Rally support and polarize identities. Subtypes: 12, 14, 16, 10, 7. Rationale: Builds loyalty through fear, pride, and tribal solidarity.
5. Confuse and Distract Goal: Overwhelm critical thinking through noise and uncertainty. Subtypes: 7, 15, 6, 17 (Red Herrings, Appeal to Probability, False Dichotomy). Rationale: Undermines clarity, trust, and consensus.

Table 8: Overview of the five high-level propaganda categories and their purposes.

Low-Level Fine-Grained Labels
1. Guilt-by-association fallacy Most frequently as nazi analogies. Includes WWII analogies, Hitler/SS references, burning people like Nazis. Could also include other reductio ad Hitlerum instances.
2. Historical Distortion / Revisionism Cherry-picking historical events, WWII nostalgia, Soviet glorification, fake quotes.
3. Dehumanization / Demonization Calling opponents subhuman, monsters, or animals; atrocity narratives.
4. Deflect and Justify I: Victimhood / Gaslighting Aggressor framed as victim, blame-shifting, moral inversion, denial of wrongdoing.
5. Deflect and Justify II: Whataboutism False moral equivalence, hypocrisy framing ("but the West invaded Iraq").
6. Accusation of Propaganda / Media Distrust Asserting "the other side lies," mocking mainstream narratives, distrust in institutions.
7. Conspiracy Narratives CIA/Nuland plots, puppet governments, Western coups, assertion-as-proof.
8. Guilt by Association Azov = entire military is Nazi; guilt through affiliation or past ties.
9. Sarcasm / Ridicule / Strawman Mocking tone, exaggeration, caricatures, eye-roll emojis.
10. Emotional Manipulation / Shock Appeal Rape/torture narratives, fearmongering, nuclear threats, mass graves.
11. False Authority / Fabrication Fake quotes, fabricated stats, unverified claims framed as facts.
12. Us vs. Them Framing / Identity Dichotomy Nationalist binaries, tribal solidarity, "with us or against us" rhetoric.
13. Realpolitik Framing / Moral Detachment Cool, analytical justification of aggression; geopolitical realism.
14. Triumphalism / Victory Framing "Russia is winning," inevitability rhetoric, "truth will prevail."
15. Information Laundering / Rumor Seeding "Unconfirmed reports," plausible deniability, sowing doubt.
16. Anti-Establishment / Anti-Globalist Framing Criticism of NATO, EU, UN, WEF; anti-West framing.
17. Logical Fallacies False dichotomy, causal oversimplification, circular reasoning, false balance, appeal to ignorance.

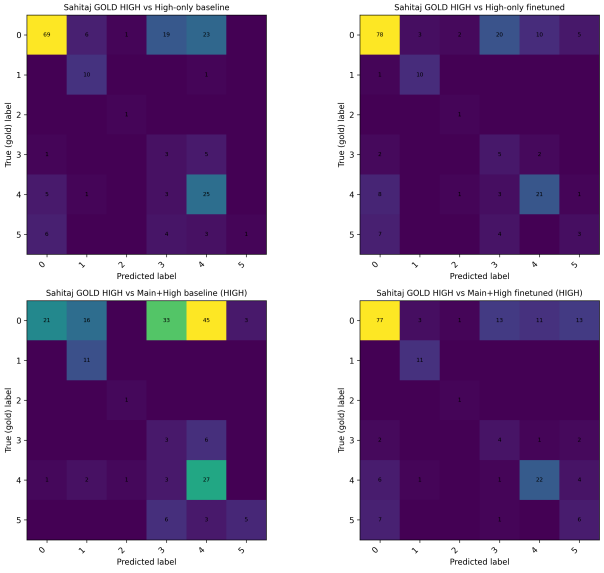
Table 9: Overview of the 17 fine-grained propaganda techniques.

E Token-level evidence for confusion patterns (example file)

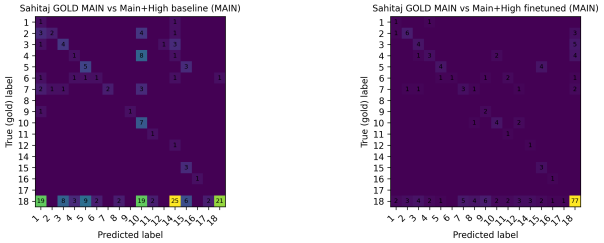
True label	Pred label	#tokens	Most frequent tokens (count)
Non-Propaganda (0)	Non-Propaganda (0)	905	ukraine (27), war (20), russia (18), putin (9), propaganda (8), russian (8), people (7), crisis (6), live (6), uk (6)
Non-Propaganda (0)	Delegitimize the Opponent (3)	567	russia (15), ukraine (11), nazi (10), russian (10), war (10), propaganda (8), people (7), state (6), nazis (5), terrorist (5)
Distort Reality (1)	Delegitimize the Opponent (3)	134	ukraine (12), russia (5), ukrainian (4), nazi (3), azov (3), battalion (3), weremember (3), neo (2), glorification (2), nazis (2)
Delegitimize (3)	Delegitimize (3)	406	nazi (25), ukraine (20), war (7), nazis (6), ukrainian (6), propaganda (6), russian (5), russia (5), support (5), azov (4)

Table 10: Illustrative excerpt of token frequency summaries for selected confusion-matrix cells (Qwen3; example file). Full token summary files for all models/settings are released in the project repository.

F Additional Confusion Matrices: Qwen3

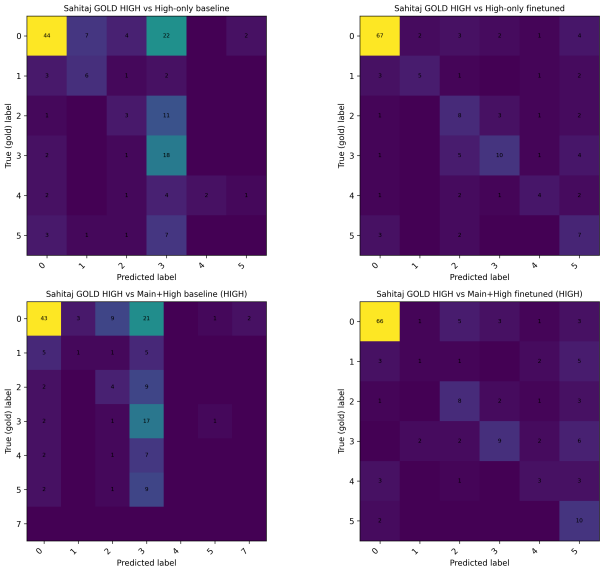


(a) HIGH

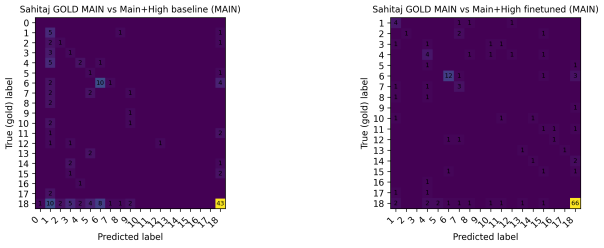


(b) MAIN

Figure 3: Qwen3 confusion matrices on the Sahitaj benchmark.



(a) HIGH



(b) MAIN

Figure 4: Qwen3 confusion matrices on our dataset.