COERCION: BENCHMARKING MORAL COERCION IN DIALOGUES VIA PSYCHOLOGICAL DIMENSION DECOUPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Moral coercion is a form of implicit manipulation that uses shared ethical norms to exert interpersonal pressure. While existing NLP models detect explicit toxicity, they struggle with moral coercion due to its polite phrasing and contextual dependency, Current datasets on moral coercion are insufficient, often relying on movie scripts that lack the pragmatic complexity of actual social interactions. To address this gap, we build COERCION, a dataset of 4,700 multi-turn dialogues sourced from social media. To model the latent manipulative mechanism of moral coercion, we design a taxonomy decomposing moral coercion into four quantifiable psychological dimensions: Obligation, Constraint, Value Judgement, and Toxicity. We further propose a two-stage dimension-aware pipeline that extracts these dimensions as structural anchors, guiding downstream Large Language Models (LLMs) toward deeper pragmatic reasoning rather than shortcut learning. Statistical analysis indicates that our dataset reflects moral coercion characteristics. Experimental results demonstrate that grounding LLMs in these dimensions improves detection accuracy. Furthermore, ablation results indicate that the joint integration of ValueJudgement and Constraint is sufficient to sustain near-optimal performance across models, demonstrating their critical contribution to the exploration of moral coercion. Our codes are available at https://anonymous.4open.science/r/BMCDPDD-88D4/

1 INTRODUCTION

Moral coercion is a pragmatic strategy in which an individual utilizes ethical concepts or elevated moral standards to exert interpersonal or emotional pressure, compelling a target to act according to specific expectations Barnhill (2014); Ienca (2017). A defining characteristic of this manipulation is the preemptive moral penalty: refusal to comply typically results in the target being labeled with negative moral judgments, such as "selfish" or "apathetic". Unlike explicit verbal toxicity, moral coercion operates through implicit structural mechanisms, primarily by leveraging public opinion or emotional appeals to restrict the target's choices, conflating voluntary personal rights with mandatory obligations, and employing double standards that demand sacrifices without reciprocal commitment. By exploiting these mechanisms, such discourse induces cognitive dissonance and systematically diminishes the target's decision-making agency Arif et al. (2024). The proliferation of digital communication platforms has amplified the reach of such manipulative discourse, necessitating robust automated detection systems.

Existing Natural Language Processing (NLP) methods effectively detect explicit toxicity and abuse. However, these methods primarily target context-free aggression and frequently fail to identify implicit, context-dependent manipulation Wiegand et al. (2021). Traditional toxicity classifiers rely heavily on lexical features, which are ineffective against the polite or altruistic syntax typically employed in moral coercion.

Despite recent advancements in safety alignment Bai et al. (2022), detecting moral coercion remains challenging due to the limitations of current Large Language Models (LLMs). First, LLMs tend to exhibit shortcut learning behaviors, relying heavily on surface-level sentiment words and specific emotional patterns rather than analyzing the logical structure of a dialogue. As a result, they misclas-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

sify negative emotions as manipulation (FP). Conversely, they fail to detect coercive requests if the speaker uses polite or harmless words (FN) Scherrer et al. (2023). Second, the decision-making process of LLMs on these tasks lacks interpretability. Direct zero-shot classification produces unstable outputs, and researchers cannot easily determine which specific conversational features influenced the model's final prediction.

Beyond model limitations, existing datasets do not support the training of robust coercion detection models. Current datasetsKu et al. (2024) rely mainly on movie scripts or other media, which fail to capture the fragmented and highly contextualized nature of implicit manipulation due to their reliance on explicit narrative structures. Furthermore, most existing datasets rely on simple binary labels and do not measure the severity of the moral coercion. Another critical issue is spurious correlations. Because current datasets contain lexical biases, models learn to associate specific topics, such as "family duties", directly with coercive labels. They memorize these topic words instead of learning the actual patterns of manipulation. Consequently, these models perform poorly when tested on new scenarios that do not contain those memorized words.

To address these challenges, our contributions are threefold. First, we constructed COERCION, a dataset of 4,700 multi-turn dialogues sourced from social media that annotates coercion intensity on a 0–5 scale to prevent models from relying on simple sentiment shortcuts. Second, we defined four quantifiable psychological dimensions including Obligation (OB), Constraint (CS), Value Judgement (VJ), and Toxicity (TX) to investigate models to capture structural manipulative patterns rather than superficial biases. Finally, we proposed a two-stage dimension-aware pipeline where an upstream LLM extracts and quantifies these dimensions to serve as structural anchors for downstream prediction, and extensive ablation studies demonstrate that optimal dimension utilization depends heavily on downstream model capacities.

2 Related Work

2.1 Implicit Discourse and Argumentation Mining

Current NLP research on implicit discourse primarily focuses on argumentation mining and propaganda detection. Argumentation mining extracts logical premises and claims from formal texts using explicit discourse markers. Propaganda detection targets rhetorical maneuvers in political campaigns Da San Martino et al. (2019). Recent studies also classify social incivility, such as condescension Perez Almendros et al. (2020) and microaggressions Breitfeller et al. (2019). However, these frameworks are not designed for moral coercion. Argumentation mining assumes rational logical exchange, whereas moral coercion relies on emotional pressure. Furthermore, models trained on political propaganda cannot detect everyday manipulation where ethical norms are used to restrict user agency.

2.2 Toxic Speech and Implicit Manipulation Detection

Existing toxicity detection models rely on lexical matching Waseem & Hovy (2016) or transformer architectures Baheti et al. (2021) to identify explicit hate speech and aggressive syntax. However, these methods fail to detect moral coercion, which masks manipulative intent using polite phrasing or societal norms. To address implicit manipulation, recent datasets like MENTALMANIP define tactics such as gaslighting based on movie scripts Ku et al. (2024). Models trained on scripted narratives often fail to generalize to complex interactions, which are highly fragmented and contextual. Following recent efforts to analyze natural user-generated content Xu et al. (2024), building datasets from social media is essential for robust manipulation detection.

2.3 Machine Ethics and Moral Psychology

In NLP, Moral Foundations Theory (MFT) Graham et al. (2013) is widely used to classify the moral framing of public discourse. To evaluate the ethical alignment of AI, datasets like ETHICS Hendrycks et al. (2021a), Social Chemistry 101 Forbes et al. (2020), and Delphi Jiang et al. (2021) test whether models can classify actions based on philosophical concepts or consensus rules. While these benchmarks evaluate static, third-person moral knowledge, they do not capture multi-turn conversational dynamics where participants use moral norms to exert pressure. Furthermore, cog-

nitive studies show that LLMs process moral scenarios using fast, System-1 heuristics Greene et al. (2001). Consequently, LLMs overreact to negative keywords and struggle to identify manipulative intent. Our dimension-aware pipeline addresses this by structuring inputs to elicit deeper, System-2 reasoning.

2.4 Benchmark Comparison

Table 1 summarizes the characteristics of existing benchmarks on machine ethics, hate speech, and mental health. Their characteristics reveal two common patterns. First, the majority of these datasets are built on isolated, single-turn statements rather than multi-turn dialogues. Second, their annotation schemes are predominantly restricted to binary or categorical labels, such as Toxic/Non-toxic, and none of them incorporate multi-level intensity scales or specific psychological dimensions.

Dataset	Domain	Dialogue	Size	Target	Intensity Scale	Psych. Dimension
Ethics VirtueHendrycks et al. (2021b)	Machine Ethics	No	33.2K	Binary (Exhibited/Not)	None	None
SemEval 2019 Task5AFortuna et al. (2019)	Hate Speech	No	13.0K	Binary (Hate/Non-hate)	None	None
TalkDownWang & Potts (2019)	Condescension	Yes	6.5K	Binary (Condescending/Not)	None	None
ToxicChatBaheti et al. (2021)	Offensiveness	Yes	2.0K	Binary (Toxic/Non-toxic)	None	None
SDCNLHaque et al. (2021)	Depression/Suicide	No	1.9K	Binary (Suicide/Non-suicide)	None	None
Implicit HateElSherief et al. (2021)	Covert Hate	No	21.5K	Categorical (Explicit/Implicit/Not)	None	None
ToxiGenHartvigsen et al. (2022)	Hate Speech	No	274.2K	Binary (Hate/Neutral)	None	None
COERCION (Ours)	**Moral Coercion**	**Yes**	**4.7K**	**Binary(Coercion/Non-Coercion)**	**6-Level (0-5)**	**O/C/V/T**

Table 1: Comparison between the proposed COERCION dataset and existing benchmarks. In the psychological dimension column, O, C, V, and T denote Obligation, Constraint, Value Judgement, and Toxicity, respectively.

3 Dataset Construction and Analysis

3.1 Taxonomy and Data Preprocessing

We propose a taxonomy that decomposes moral coercion into dialogue-level metrics and psychological dimensions. As shown in Table 1, the COERCION dataset provides two global annotations per interaction: a binary detection label indicating the presence (1) or absence (0) of moral coercion, and an overall intensity ordinal score from 0 to 5 measuring the severity of the pressure. We define four psychological dimensions to model coercion mechanisms, and we incorporate standard toxicity datasets into our evaluation to test cross-domain generalization and distinguish implicit coercion from explicit toxicity.

We constructed the candidate pool by collecting interactions from major social media platforms via a keyword-driven web crawler. To prevent data leakage, we first partitioned the dialogue data into training, validation, and test sets. Given the scarcity of rare edge-case moral coercion in the raw text, we applied an iterative process of data generation and manual filtering independently within each split. In each round, GPT-4o generated implicit coercive examples and non-coercive counterfactual pairs. Expert annotators then reviewed these candidate texts to remove duplicates, incomplete contexts, and inaccurate generations before advancing the finalized high-quality data to the formal labeling phase. The partition yields training, validation, and test sets with 3,432, 634, and 634 instances, respectively.

3.2 Human Annotation

All dialogue interactions were collected from public social media platforms in strict compliance with their respective terms of service for research purposes. To eliminate privacy risks and ensure complete user anonymity, we implemented a rigorous data sanitization protocol that automatically and manually removed all personally identifiable information (PII), including usernames, user IDs, hyperlinked mentions, and specific real-world entity references, leaving only de-identified textual contexts for analysis.

We established a human annotation pipeline with a specialized panel of six expert annotators (four males and two females) holding educational backgrounds in Psychology, Linguistics, or Computer Science and possessing a profound understanding of the relevant cultural background. Prior to the

formal annotation phase, the annotators participated in specialized training sessions to extensively review the annotation guidelines, which included clear definitions, boundary cases, and standardized examples of moral coercion. To maximize annotation quality and mitigate subjectivity, each text instance was independently evaluated by three annotators under a gender-balanced assignment strategy of two males and one female per task to neutralize potential gender-based perceptual biases regarding emotional pressure.

For each instance, the annotators provided a binary judgment on the presence of moral coercion and an ordinal score from 0 to 5 for coercion intensity. We employed a majority voting strategy for the binary classification task, where the majority consensus determined the final label. For instances identified as moral coercion, the intensity score was calculated by averaging the affirmative ratings, which was then subjected to a review process to be rounded to the nearest discrete intensity level on our scale. Following a quality review process to filter incomplete submissions and resolve extreme outliers, the finalized dataset comprises 4,700 high-quality annotated instances. We measured the inter-annotator agreement using Fleiss' Kappa based on the binary judgments, yielding a score of 0.623 that indicates a substantial agreement among the annotators and demonstrates the effectiveness of the standardized guidelines.

3.3 Micro-Level: Quantifiable Psychological Dimensions

To unpack the underlying mechanisms driving the macro-level intensity, we evaluate each dialogue across four distinct, theoretically grounded psychological dimensions. Each dimension is scored on a continuous scale from 0.0 to 1.0, categorized into four discrete tiers: None (0.0), Low (0.1–0.3), Moderate (0.4–0.7), and High (0.8–1.0).

Obligation (OB) Drawing upon deontological theories, OB measures the enforcement of perceived moral duties and social contracts, escalating from gentle appeals to general norms (Low), to explicit pressure invoking "musts" (Moderate), and ultimately to unavoidable demands framed as absolute moral imperatives (High)Cushman (2013).

Constraint (CS) Grounded in philosophical definitions of coercion and pragmatic theories of cognitive manipulation, CS evaluates the restriction of the listener's agency, progressing from a subtle narrowing of acceptable options (Low), to explicit logical traps and conditional threats (Moderate), and peaking at the total deprivation of choice and psychological cornering (High)Nozick (1969); Maillat & Oswald (2009).

Value Judgement (VJ) Because moral evaluations are inherently linked to power dynamics and psychological contamination, VJ assesses the severity of moral shaming and guilt induction, ranging from mild subjective preferences (Low), to loaded moral terms targeting character (Moderate), up to severe moral condemnation (High)Tangney et al. (2007); Haidt (2003).

Toxicity (TX) To cleanly separate coercive strategies from overt hostility, TX quantifies the explicit aggressiveness of the tone. Aligning with standard toxic language detection, this metric spans from passive-aggressive sarcasm (Low) to overtly insulting remarks (Moderate) and hateful abuse (High)Dixon et al. (2018); Fortuna & Nunes (2018).

3.4 Two-Stage Dimension-Aware Pipeline

We propose a two-stage dimension-aware pipeline that explicitly decouples dimension extraction from the final classification. As shown in Figure 1,

Stage 1: Dimension Extraction Given a dialogue $D = \{u_1, u_2, \ldots, u_n\}$ consisting of n utterances, we employ an upstream model M_{up} to extract the four psychological dimensions: Obligation (OB), Constraint (CS), Value Judgement (VJ), and Toxicity (TX). The extraction produces a psychological dimension profile $F = \{F_{global}, F_{turn}\}$:

- **Turn-level Analysis** (F_{turn}): For each utterance u_i, the model assigns a discrete categorical level (None, Low, Moderate, High) and a continuous intensity score $s \in [0, 1]$ to all four psychological dimensions.

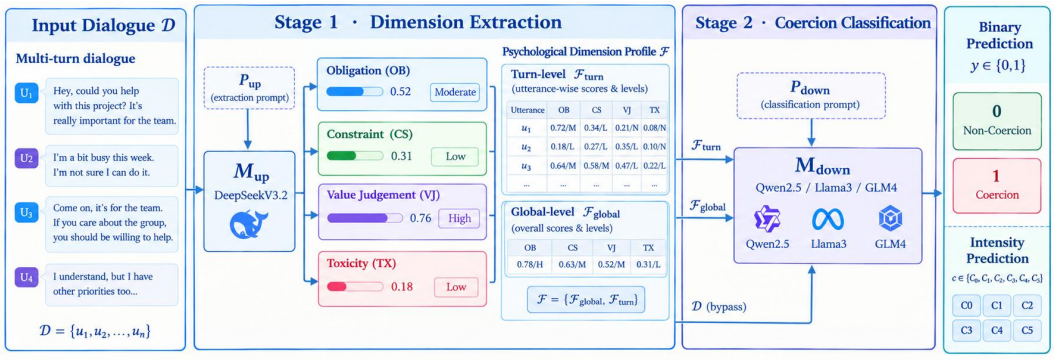


Figure 1: Two-stage dimension-aware pipline for moral coercion dection.

- **Global-level Analysis** (F_{global}): The model assigns an overall level and score for each psychological dimension across the entire dialogue D. Rather than computing a simple mathematical average, this global evaluation accounts for peak severity, accumulated pressure, and mitigating repairs.

Stage 2: Coercion Classification In the second stage, the extracted dimension profile F is formatted into a structured text representation and concatenated with the original dialogue D. A downstream model M_{down} utilizes this augmented context to make the final binary prediction and intensity prediction. Formally, the complete pipeline is defined as:

$$F = M_{up}(P_{up} \oplus D)$$
$$y = M_{down}(P_{down} \oplus F \oplus D)$$

where $\oplus$ denotes string concatenation, P_{up} and P_{down} represent the task-specific prompts for extraction and classification respectively, and $y \in \{0, 1\}$ is the final prediction (0 for non-coercion, 1 for coercion) and $c \in \{c_0, c_1, \ldots, c_5\}$ is the intensity prediction. The explicit separation provides M_{down} with intermediate reasoning anchors, allowing it to focus strictly on the decision boundary.

3.5 Ethics Statement

All dialogue interactions were collected from public social media platforms in strict compliance with their respective terms of service for research purposes. To eliminate privacy risks and ensure complete user anonymity, we implemented a rigorous data sanitization protocol that automatically and manually removed all personally identifiable information (PII), including usernames, user IDs, hyperlinked mentions, and specific real-world entity references, leaving only de-identified textual contexts for analysis. Given the sensitive nature of moral coercion, we prioritized annotator well-being by enforcing fair compensation, reasonable daily workload limits, and periodic check-ins; all annotators maintained a healthy and normal psychological state throughout the annotation phase. To prevent potential malicious misuse or improper deployment of coercive discourse patterns, the dataset is not distributed openly without oversight; access is strictly granted to academic researchers upon formal evaluation and approval of their application, subject to a signed Data Use Agreement.

4 Dataset Statistics

To accurately profile how moral coercion manifests in interactions and verify the validity of Coercion, we evaluate the dataset across four dimensions: persuasion techniques and moral foundations in the main text, alongside emotional distributions and t-SNE semantic space in Appendix A.3 and Appendix A.4. By systematically evaluating the specific rhetorical strategies, ethical pressures, and varied emotional states present in the text, we demonstrate that coercion relies on complex implicit manipulation, making it inherently difficult for traditional lexical methods to detect. And the severe structural overlap observed in the t-Distributed Stochastic NeighborEmbedding(t-SNE) semantic space suggests that coercive and non-coercive dialogues are highly entangled, underscoring the critical necessity of our specialized dataset for training robust models.

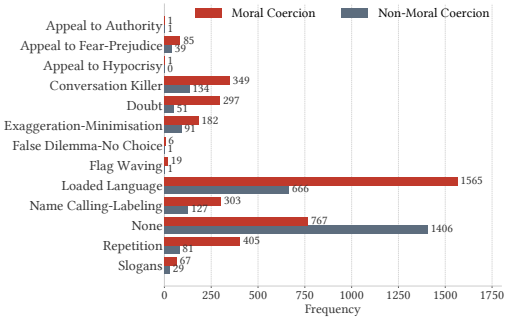


Figure 2: Frequency distribution of manipulation techniques across coercive and non-coercive dialogues.

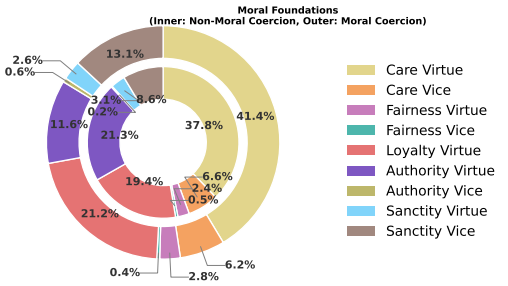


Figure 3: Dialogue-level comparison of moral foundations. The chart illustrates how different moral virtues and vices map onto coercive and non-coercive interactions.

4.1 Persuasion Techniques

To automatically identify persuasion strategies within the dialogues, we utilized a multilingual sequence classification model DaniilOr (2024) fine-tuned on the SemEval 2023 Task 3 dataset Piskorski et al. (2023).

Specifically, the evaluation taxonomy consists of 12 standard persuasive techniques alongside a baseline category: Appeal to Authority, Appeal to Fear-Prejudice, Appeal to Hypocrisy, Conversation Killer, Doubt, Exaggeration-Minimisation, False Dilemma-No Choice, Flag Waving, Loaded Language, Name Calling-Labeling, Repetition, Slogans, and None (detailed definitions are provided in Appendix A.1).

As shown in Figure 2, there is a stark contrast in technique distribution between the two classes. The results indicate that Loaded Language is the most dominant technique in moral coercion, significantly outnumbering its presence in non-coercive dialogues. Furthermore, aggressive and suppressive tactics such as Repetition, Conversation Killer, and Name Calling-Labeling are heavily skewed towards the coercive class. The results align with the characteristics of moral coercion, where manipulators frequently use emotionally charged and absolute language to corner the victim.

4.2 Moral Foundations

To uncover the underlying ethical frameworks exploited in coercive discourse, we adopted the Moral Foundations Theory Graham et al. (2013). The theory posits that human morality is driven by five universal foundations, each expressed as a bipolar virtue-vice scale, resulting in ten distinct label dimensions (detailed descriptions are provided in Appendix A.2).

We employed a classification model based on Moral Foundations Theory to categorize the dialogues into the ten dimensions. As illustrated in Figure 3, coercive dialogues invoke moral dimensions at a significantly higher frequency than non-coercive dialogues. The *care_virtue* dimension is the most prominent, indicating that manipulators frequently frame coercive requests as compassion or acting in the target's best interest. Furthermore, dimensions enforcing social binding and hierarchical compliance (e.g., *loyalty_virtue* and *authority_virtue*) are heavily utilized, reflecting general norms of in-group conformity. Notably, the *sanctity_vice* dimension exhibits a significant increase, indicating that moral coercion systematically relies on moral shaming and negative labeling to enforce compliance. The co-occurrence of different dimensions (e.g., *Care* and *Authority*) demonstrates the complex interplay of moral frameworks in interactions. This structural complexity distinguishes authentic coercive discourse from the explicit conflicts typically found in scripted datasets.

5 EXPERIMENTS

5.1 EXPERIMENT SETTING

To evaluate the COERCION dataset and benchmark state-of-the-art models on recognizing moral coercion, we formulated two core experimental tasks: Moral Coercion Detection, a binary classification problem to identify the presence of coercion, and Moral Coercion Intensity Detection, designed to predict the severity of the applied pressure.

We analyzed the performance of representative open-weights large language models, specifically GLM-4-9B-Chat, Qwen2.5-14B-Instruct and Llama-2-13B-Chat, alongside a robust traditional pre-trained language model, RoBERTa-base. All reported metrics representing the mean of three independent runs.

5.2 MORAL COERCION DETECTION

We frame moral coercion detection as a binary classification task and evaluate three configurations: cross-domain and in-domain evaluation, and in-context learning. Performance is measured using Accuracy (Acc), Precision (Prec), Recall (Rec), and the F1 score.

Model	**Acc**	**Prec**	**Rec**	**F1**
DeBERTA-v3-EmoBandel (2022)	0.7208	0.7192	0.7667	0.7421
MoralFound-ClfArdag (2024)	0.6736	0.7004	0.6667	0.6816
RoBERTa-GoEmoKumar (2025)	0.7487	0.7500	0.7800	0.7647
RoBERTa-MoralKim et al. (2024)	0.7173	0.7344	0.7200	0.7270

Table 2: Baseline performance on emotion and moral classification.

Training Dataset	**GLM-4-9B-Chat**				**Qwen2.5-14B-Instruct**				**Llama-2-13B-Chat**				**RoBERTa-base**			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Ethics Virtue	0.491	0.518	0.401	0.452	0.738	0.717	0.825	0.768	0.486	0.615	0.048	0.089	0.471	0.474	0.081	0.139
SemEval 2019 Task5A	0.492	0.999	0.030	0.059	0.700	0.735	0.669	0.700	0.480	0.999	0.006	0.012	0.494	0.790	0.045	0.086
TalkDown	0.492	0.999	0.030	0.059	0.505	0.950	0.057	0.108	0.476	0.000	0.000	0.000	0.522	0.746	0.133	0.225
ToxicChat	0.559	0.947	0.163	0.278	0.691	0.754	0.608	0.673	0.492	0.778	0.042	0.080	0.494	0.824	0.042	0.080
SDCNL	0.505	0.950	0.057	0.108	0.686	0.629	0.976	0.765	0.478	0.999	0.003	0.006	0.522	0.529	0.807	0.639
Implicit Hate	0.560	0.544	0.991	0.702	0.669	0.614	0.990	0.758	0.662	0.634	0.843	0.724	0.524	0.524	0.999	0.687
ToxiGen	0.710	0.699	0.783	0.739	0.596	0.565	0.997	0.721	0.655	0.604	0.991	0.750	0.721	0.688	0.855	0.762
COERCION (Ours)	**0.842**	**0.847**	**0.852**	**0.850**	**0.834**	**0.847**	**0.834**	**0.841**	**0.831**	**0.832**	**0.849**	**0.841**	**0.838**	**0.823**	**0.880**	**0.850**

Table 3: Performance of Models Fine-tuned on Various Related Datasets vs. the Proposed **COERCION** Dataset. Acc, Prec, Rec, and F1 stand for accuracy, binary precision, binary recall, and macro F1 respectively.

Cross-Domain and In-Domain Evaluation To verify whether moral coercion is distinctly different from general toxicity or standard emotions, we evaluate several pre-trained domain-specific models directly on COERCION. As detailed in Table 2, the F1 scores of the models fine-tuned for emotion detection or moral foundations range from 0.6816 to 0.7421. Similarly, we conduct coercion detection by fine-tuning LLMs and RoBERTa on existing abusive language, machine ethics, and mental health datasets. As shown in Table 3, the performance remains poor when tested on COERCION; for instance, Llama-2-13B-Chat fine-tuned on ToxicChat yields an F1 of 0.080, while training on Implicit Hate achieves an F1 of 0.724. These results indicate that moral coercion possesses unique implicit semantic patterns are not effectively captured by models trained solely on explicit toxicity corpora. Conversely, to align the models with the specific task, we fine-tune them directly on the COERCION training partition. As presented in Table 3, this in-domain fine-tuning yields substantial performance enhancements across all models: GLM-4-9B-Chat and RoBERTa-base achieve an F1 score of 0.850, while Qwen2.5-14B-Instruct and Llama-2-13B-Chat reach 0.841. This demonstrates that direct parameter updates on our dataset effectively teach the models to recognize moral coercion, significantly outperforming zero-shot and cross-domain baselines.

5.3 MORAL COERCION INTENSITY PREDICTION

The purpose of the experiment was to evaluate the zero-shot capabilities of LLMs on the fine-grained task, and to assess the performance gains enabled by fine-tuning on COERCION. We formulated a 6-class intensity prediction.

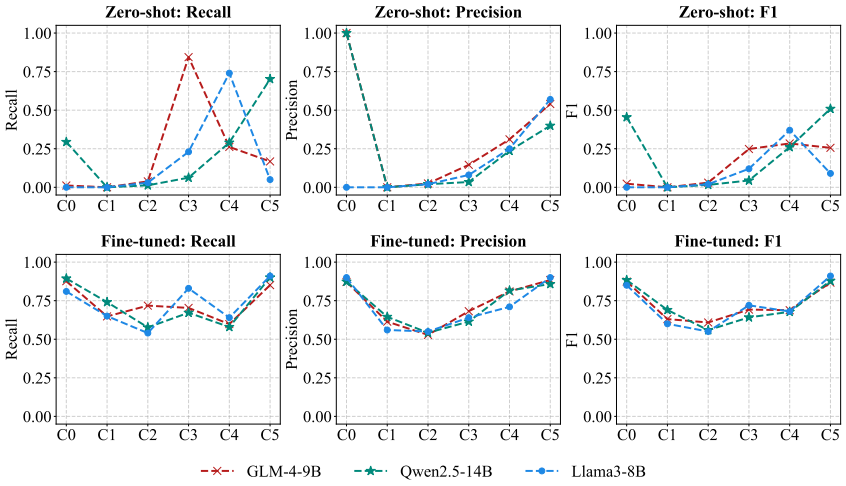


Figure 4: Performance profiles across all moral coercion intensity classes.

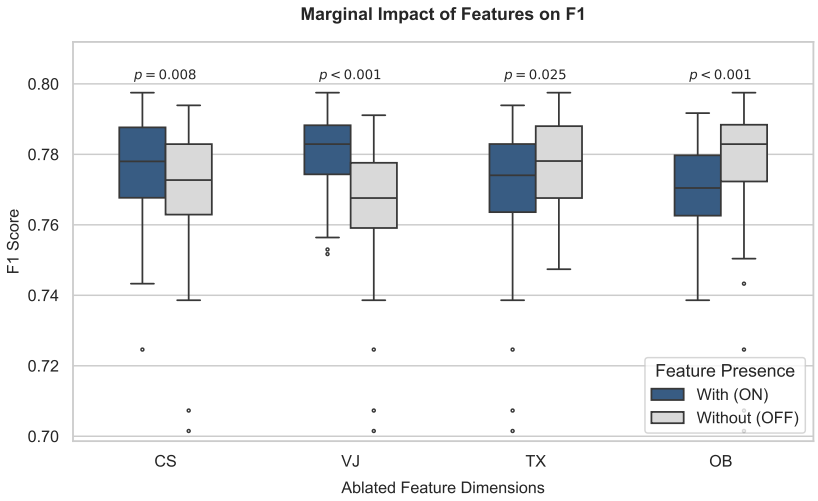


Figure 5: Feature marginal impact

Zero-Shot vs. Fine-Tuning Performance Figure 4 presents the model performance on the 6-class intensity prediction task, , where c_0 denotes the lowest intensity and c_5 denotes the highest intensity. Under the zero-shot setting, all evaluated models performed poorly. Conversely, fine-tuning significantly improved performance across all metrics. These results indicate that while zero-shot LLMs fail to accurately classify coercion intensity, supervised fine-tuning enables them to effectively distinguish fine-grained severity levels.

5.4 Integration Granularity and Dimension Ablation Study

To instantiate our proposed two-stage dimension-aware pipeline, we utilized DeepSeek-V3.2 as upstream pragmatic parsers ($\mathcal{M}_{up}$). This comparative selection allowed us to objectively observe how different inductive biases during dimension extraction affected downstream dependency. For downstream classification ($\mathcal{M}_{down}$), we evaluated four LLMs with parameter scales ranging from 7B to 14B to comprehensively validate the generalizability of the extracted psychological dimensions.

6 Ablation Study and Feature Dynamics

To evaluate the impact and necessity of the extracted pragmatic dimensions, we analyze the marginal contribution of each dimension using paired statistical significance tests in Figure 5 and conduct an ablation study across four backbones in Table 4.

Control Mode	Qwen2.5-14B		Qwen2.5-7B		GLM-4-9B		Llama-3-8B	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Vanilla Baseline	0.672	0.727	0.707	0.757	0.683	0.759	0.640	0.744
− Feature Semantics	0.726	0.713	**0.759**	0.773	0.685	0.765	0.650	0.748
− Feature Values	0.634	0.705	0.702	0.757	0.527	0.689	0.524	0.687
+ All Features (Ours)	**0.756**	<u>0.789</u>	0.738	<u>0.785</u>	**0.726**	<u>0.786</u>	<u>0.672</u>	<u>0.758</u>
+ Subset Features (CS+VJ)	<u>0.743</u>	**0.796**	<u>0.753</u>	**0.796**	<u>0.724</u>	**0.786**	**0.705**	**0.776**

Table 4: Ablation study across backbones. **Bold** and <u>underlined</u> denote best and second-best results.

Marginal Feature Contributions To isolate the independent impact of each dimension, Figure 5 compares F1 distributions under presence (*With*) versus absence (*Without*) conditions across all trial configurations. Paired significance tests reveal contrasting marginal effects across dimensions. Specifically, *ValueJudgement* (VJ) and *Constraint* (CS) both exhibit statistically significant positive marginal contributions to performance. Conversely, incorporating *Toxicity* (TX) or *Obligation* (OB) leads to statistically significant performance reductions, demonstrating that accumulating all available features does not monotonically improve classification outcomes.

These empirical observations align with the linguistic structure of moral coercion in text. Coercive discourse infrequently uses overt hostility (*Toxicity*) or explicit directives (*Obligation*, e.g., "You must obey me"). Instead, it relies predominantly on moral evaluation (*ValueJudgement*, e.g., "A responsible person wouldn't refuse") combined with restrictions on acceptable behavior (*Constraint*).

Feature Ablation As presented in Table 4, incorporating the full feature set (**+ All Features**) consistently outperforms the *Vanilla Baseline* across all architectures. To verify whether downstream models utilize the numerical intensity values rather than prompt layout artifacts, we evaluate two control conditions:

- **Feature Values**: replacing extracted intensity scores with random values. This setup causes a substantial performance decline across all backbones, dropping F1 scores below the baseline. It indicates that classification accuracy relies on the numerical fidelity provided by the upstream extraction.
- **Feature Semantics**: masking feature names with abstract tokens. This setup maintains competitive performance, indicating that models rely primarily on numerical intensity patterns rather than explicit feature labels.

7 Conclusion

In this paper, we investigated the detection of moral coercion in digital discourse. To support the task, we built **COERCION**, a dataset comprising 4,700 multi-turn dialogues. Based on moral psychology, we formulate four quantifiable psychological dimensions: *Obligation* (OB), *Constraint* (CS), *ValueJudgement* (VJ), and *Toxicity* (TX) to systematically unravel the underlying mechanics of moral coercion. Statistical analysis indicates that COERCION reflect the characteristics of moral coercion. Furthermore, COERCION and its psychological dimensions framework provide a reliable foundation for training and evaluating downstream models on implicit manipulation tasks.

To improve LLMs' ability to recognize moral coercion, we proposed a two-stage dimension-aware pipeline. Our evaluations demonstrated that explicit dimension integration significantly outperforms zero-shot baselines. Furthermore, in-domain fine-tuning effectively corrects the substantial prediction errors observed in base models, enabling them to accurately classify fine-grained severity levels.

Crucially, ablation experiments reveal how these dimensions interact, establishing *ValueJudgement* (VJ) and *Constraint* (CS) as the primary drivers of moral coercion detection. In contrast, the limited predictive utility of *Obligation* (OB) underscores a key nuance: moral coercion hinges on subtle, implicit evaluations rather than overt explicit mandates.

References

M. Murat Ardag. Moral foundations classifier. https://doi.org/10.57967/hf/2774, 2024. Accessed: 2026-03-10.

Aahan Arif, Muskaan Abdul Qadir, Russell Seth Martins, et al. The impact of cyberbullying on mental health outcomes amongst university students: A systematic review. *PLOS Mental Health*, 2024.

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4846–4862, 2021.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Elron Bandel. Deberta-v3-large emotion. `https://huggingface.co/Elron/deberta-v3-large-emotion`, 2022.

Anne Barnhill. What is manipulation? In *Manipulation: Theory and Practice*, pp. 51–72. Oxford University Press, 2014.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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)*, pp. 1664–1674, 2019.

Fiery Cushman. Action, outcome, and value in dual-system framework for morality. *Personality and social psychology review*, 17(3):273–292, 2013.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 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)*, pp. 5636–5646, 2019.

DaniilOr. multilingual_persuasion_techniques. `https://huggingface.co/DaniilOr/multilingual_persuasion_techniques`, 2024.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67–73, 2018.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 345–363. Association for Computational Linguistics, 2021.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 653–670, 2020.

Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

Paula Fortuna, Juan Soler-Company, and Sérgio Nunes. Stop PropagHate at SemEval-2019 Tasks 5 and 6: Are abusive language classification results reproducible? In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M." Mohammad (eds.), *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 745–752". Association for Computational Linguistics, 2019.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47:55–130, 2013.

J D Greene, R B Sommerville, L E Nystrom, J M Darley, and J D Cohen. An fmri investigation of emotional engagement in moral judgment. *Science*, 293(5537):2105–2108, 2001.

Jonathan Haidt. The moral emotions. In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (eds.), *Handbook of affective sciences*, pp. 852–870. Oxford University Press, 2003.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. Deep learning for suicide and depression identification with unsupervised label correction, 2021.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for implicit and adversarial hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.

Marcello Ienca. The right to cognitive liberty. *Scientific American*, 317(2), 2017.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, et al. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*, 2021.

Jaehong Kim, Chaeyoon Jeong, Seongchan Park, Meeyoung Cha, and Wonjae Lee. How do moral emotions shape political participation? a cross-cultural analysis of online petitions using language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 16274–16289. Association for Computational Linguistics, 2024.

Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Mentalmanip: A dataset for fine-grained analysis of mental manipulation in conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Lakshya Kumar. Roberta-large goemotions (optimized thresholds and focal loss). `https://huggingface.co/Laksssshya/roberta-large-goemotions`, 2025.

Sam Lowe. roberta-base-go_emotions. `https://huggingface.co/SamLowe/roberta-base-go_emotions`, 2023.

Didier Maillat and Steve Oswald. Defining manipulative discourse: The pragmatics of cognitive illusions. *International Review of Pragmatics*, 1(2):348–370, 2009.

Robert Nozick. Coercion. In Sidney Morgenbesser, Patrick Suppes, and Morton White (eds.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, pp. 440–472. St. Martin's Press, 1969.

Carla Perez Almendros, Luis Espinosa Anke, and Steven Schockaert. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5891–5902, 2020.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 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)*, pp. 2343–2361. Association for Computational Linguistics, 2023.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. Evaluating the moral beliefs encoded in llms. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 3747—-3764, 2023.

June P Tangney, Jeffrey Stuewig, and Debra Mashek. Moral emotions and moral behavior. *Annual review of psychology*, 58:345–372, 2007.

Zijian Wang and Christopher Potts. Talkdown: A corpus for condescension detection in context. 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)*, pp. 3711–3719, 2019.

Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. Implicitly abusive language–what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 576–587. Association for Computational Linguistics, 2021.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1), 2024.

A Appendix

A.1 Taxonomy Definitions

The detailed definitions of the 12 standard persuasive techniques and the baseline category used in our evaluation taxonomy are as follows:

Appeal to Authority: Stating that a particular entity considered as an authority is the source of the information to give weight to an argument.

Appeal to Fear-Prejudice: Promoting or rejecting an idea through the repulsion or fear of the audience towards it.

Appeal to Hypocrisy: Attacking the target by charging them with hypocrisy or inconsistency to undermine their reputation.

Conversation Killer: Words or phrases that discourage critical thought and meaningful discussion about a given topic.

Doubt: Questioning personal attributes or character to challenge general credibility or quality.

Exaggeration-Minimisation: Representing something in an excessive manner or making it seem less important than it really is.

False Dilemma-No Choice: A logical fallacy presenting only two options or sides when multiple choices exist, or instructing specific actions to eliminate alternatives.

Flag Waving: Justifying an idea by exalting group pride or highlighting localized benefits.

Loaded Language: Utilizing words and phrases with strong emotional implications to influence and convince the audience.

Name Calling-Labeling: Directing loaded or insulting labels at an individual, group, object, or activity.

Repetition: Using the same phrase repeatedly to persuade the audience.

Slogans: A brief and striking phrase that acts as an emotional appeal and may incorporate stereotyping.

None: The baseline category indicating the absence of any defined manipulative rhetorical technique.

A.2 Moral Foundations Dimensions

The Moral Foundations Theory Graham et al. (2013) posits that human morality is driven by five universal foundations, each expressed as a bipolar virtue-vice scale, resulting in the following ten distinct label dimensions used in our evaluation:

Care Virtue / Care Vice: Compassion, kindness, and nurturing versus harm, cruelty, and suffering.

Fairness Virtue / Fairness Vice: Justice, equality, and reciprocity versus cheating, inequality, and injustice.

Loyalty Virtue / Loyalty Vice: Loyalty, patriotism, and self-sacrifice versus betrayal, treason, and disloyalty.

Authority Virtue / Authority Vice: Respect, tradition, and order versus subversion, disobedience, and chaos.

Sanctity Virtue / Sanctity Vice: Purity, sanctity, and nobility versus degradation, contamination, and impurity.

A.3 Emotion Distribution

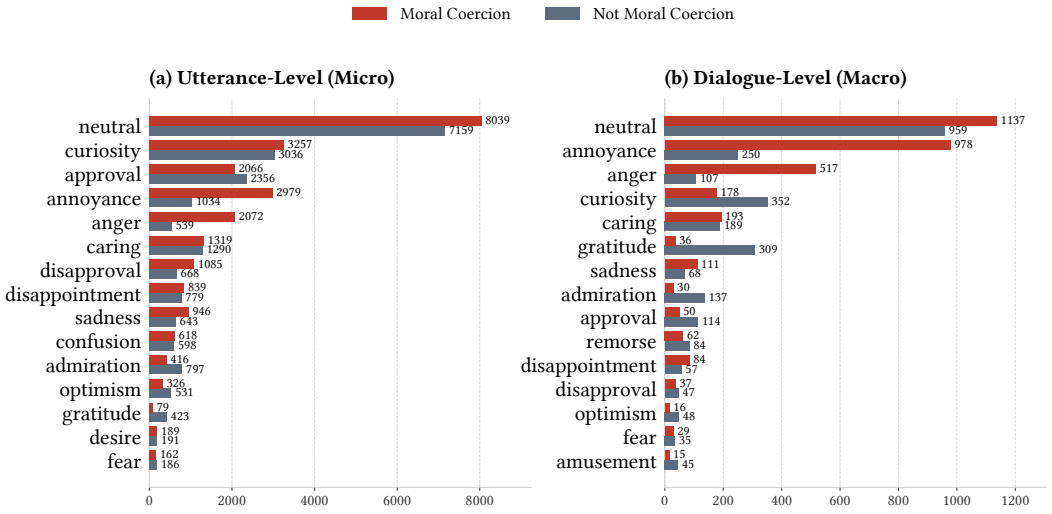


Figure 6: Comparative analysis of emotional distributions. Subfigure (A) illustrates the micro-level emotions per utterance, while Subfigure (B) depicts the macro-level overarching emotions per dialogue, highlighting the notable increase of 'annoyance' and 'anger' in coercive contexts.

To analyze the emotions in moral coercion, we employed a RoBERTa-base model fine-tuned on the GoEmotions dataset Lowe (2023) to recognize the dominant emotions at both the utterance and dialogue levels. As shown in Figure 6, while "neutral" serves as a common baseline across both categories, the distributions of specific emotions differ significantly. As illustrated in the right panel of Figure 6, dialogue-level analysis indicates that moral coercion exhibits significantly higher frequencies of annoyance and anger. In contrast, non-coercive dialogues display higher proportions of positive emotions, such as curiosity, caring, and gratitude. The combination of elevated annoyance and anger within generally neutral conversations reflects the implicit nature of moral coercion. This emotional distribution differs from scripted datasets, which typically exhibit more direct and explicit hostility for dramatic effect.

A.4 t-Distributed Stochastic Neighbor Embedding

We evaluated the separability of coercive and non-coercive texts in the semantic embedding space. We extracted sentence embeddings for the entire dataset using the all-MiniLM-L12-v2 model and projected them into a 2D space via t-SNE, as shown in Figure 7. The visualization reveals a substantial overlap between coercive (red) and non-coercive (grey) instances, without distinct clustering boundary. The substantial overlap demonstrates our assertion regarding the difficulty of identifying moral coercion: most coercive texts share identical vocabulary and syntactic structures with normal conversations. It suggest that simple keyword-based heuristics or shallow semantic embeddings are insufficient for recognizing moral coercion, which indicates the necessity of our high-quality, context-aware dataset for training more advanced language models.

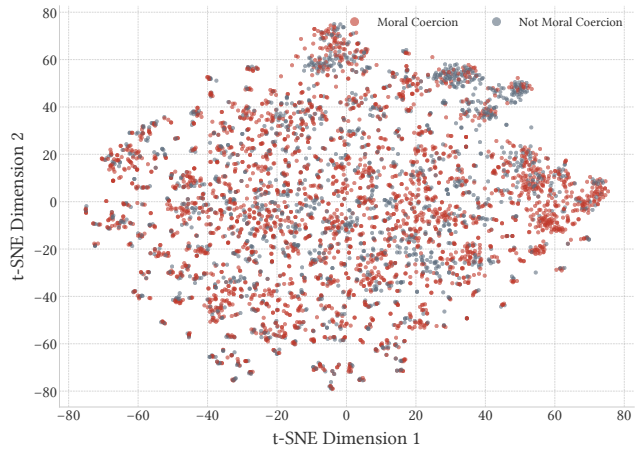


Figure 7: t-SNE visualization of the COERCION dataset using *all-MiniLM-L12-v2* embeddings.

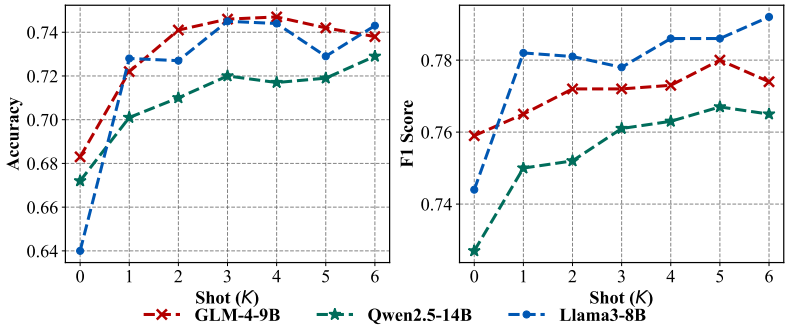


Figure 8: Few-shot Evaluation Results of LLMs

A.5 Moral Coercion Detection

In-Context Learning (Few-Shot) We evaluated the impact of providing K-shot examples ($K \in [0, 6]$), with detailed results presented in Figure 8, , where the base instructions are augmented with randomly selected in-context demonstrations, and each K-shot configuration incorporates K coercive and K non-coercive examples to total K exemplars for maintaining a balanced class distribution. The models exhibited similar performance trends as the number of examples increased.

A.6 Moral Coercion Intensity Prediction

The purpose of the experiment was to evaluate the zero-shot capabilities of LLMs on the fine-grained task, and to assess the performance gains enabled by fine-tuning on COERCION. We formulated a 6-class intensity prediction.

Zero-Shot vs. Fine-Tuning Performance Figure 4 presents the model performance on the 6-class intensity prediction task. Under the zero-shot setting, all evaluated models performed poorly, particularly on minority classes. Conversely, fine-tuning significantly improved performance across all metrics. These results indicate that while zero-shot LLMs fail to accurately classify coercion intensity, supervised fine-tuning enables them to effectively distinguish fine-grained severity levels.

Ordinal Metrics and Error Distance Analysis Because moral coercion intensity is naturally ordinal, standard accuracy treats minor boundary shifts identically to severe misclassifications. To capture degree-sensitive alignment, we incorporate Quadratic Weighted Kappa (QWK) and adjacent accuracy (Acc $\pm$ 1), alongside the error distance distribution ($\Delta = |y - \hat{y}|$), as detailed in Table 5.

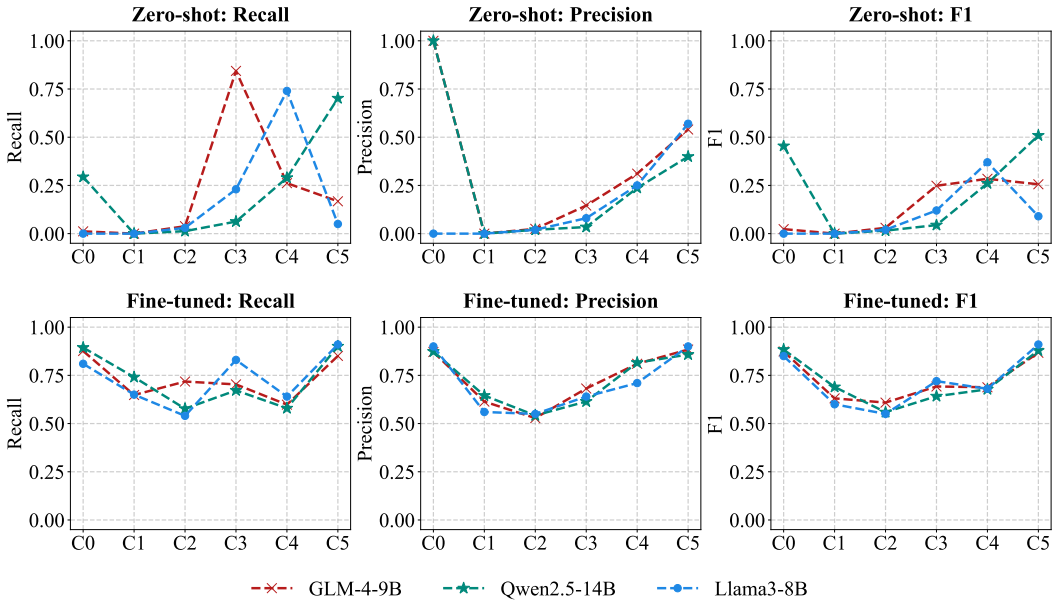


Figure 9: Performance profiles across all coercion severity classes (C0 to C5) and the Macro average.

Zero-shot LLMs display a poor capacity for ordinal scaling. Even the strongest zero-shot baseline (Qwen2.5) achieves a QWK below 0.47, with significant error mass dispersed across larger distances ($\Delta \geq 2$). Task-specific fine-tuning fundamentally restructures these error distributions across all architectures: QWK scores double to over 0.80 (peaking at 0.8298 for Qwen2.5), and exact match rates ($\Delta = 0$) jump from under 32% to around 76%. Crucially, all fine-tuned models restrict over 88% of predictions to within a single step of the ground truth (Acc $\pm 1 \geq 0.8817$). This dramatic error concentration demonstrates that fine-tuning successfully confines misclassifications to adjacent severity levels, aligning model behavior closely with human consensus.

Table 5: Performance on Severity Rating: Comparison of Zero-shot and Fine-tuned Models.

Metric	Zero-shot			Fine-tuned		
	GLM-4-9B-Chat	Qwen2.5-14B-Instruct	Llama-3-8B-Instruct	GLM-4-9B-Chat	Qwen2.5-14B-Instruct	Llama-3-8B-Instruct
Overall Metrics						
Strict Acc	0.1798	0.3139	0.1640	0.7666	0.7681	0.7618
Acc ±1	0.4968	0.5315	0.5347	0.8817	0.8833	0.8833
QWK	0.3399	0.4685	0.3401	0.8215	0.8298	0.8030
Error Distance Distribution (%)						
$\Delta = 0$	17.98	31.39	16.40	76.66	76.81	76.18
$\Delta = 1$	31.70	21.77	37.07	11.51	11.51	12.15
$\Delta = 2$	37.70	24.92	33.60	4.73	4.73	3.47
$\Delta = 3$	10.25	14.67	5.99	2.05	2.21	2.05
$\Delta = 4$	1.74	3.94	6.62	4.26	3.94	5.36
$\Delta = 5$	0.63	3.31	0.32	0.79	0.79	0.79

A.7 Integration Granularity and Dimension Ablation Study

To instantiate our proposed two-stage dimension-aware pipeline, we utilized DeepSeek-V3.2 as upstream pragmatic parsers ($\mathcal{M}_{up}$). This comparative selection allowed us to objectively observe how different inductive biases during dimension extraction affected downstream dependency. For downstream classification ($\mathcal{M}_{down}$), we evaluated four LLMs with parameter scales ranging from 7B to 32B to comprehensively validate the generalizability of the extracted psychological dimensions. To determine the optimal context window, we experimented with three granularity modes: Global (holistic dialogue context), Turn (utterance-level context), and Both (concatenation of both).

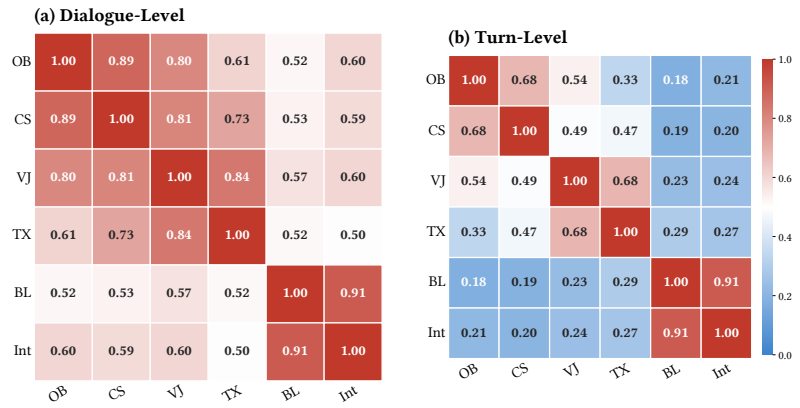


Figure 10: Pearson correlation heatmaps between the psychological dimensions and the target tasks extracted by DeepSeek-V3.2, where BL and Int denote Binary_Label and Intensity, respectively. (a) Dialogue-level correlations. (b) Turn-level correlations.

Dimension Correlation To investigate the effectiveness of the proposed dimensions, we analyzed the Pearson correlation between the four extracted dimensions (OB, CS, VJ, TX) and the annotated labels (Binary_Label, Intensity), with key dialogue-level and turn-level correlations plotted in Figure 10. At the dialogue level, all four dimensions exhibited positive correlations with both target tasks; specifically, the correlations with Intensity ranged from 0.50 (TX) to 0.60 (OB and VJ), while the correlations with Binary_Label ranged from 0.52 (OB and TX) to 0.57 (VJ). Conversely, at the turn level, these correlations decreased significantly, dropping to a range of 0.18 to 0.29 across all dimension-label pairs. Concurrently, comparative experiments utilizing GPT-4o for dimension extraction yielded consistent correlation patterns, further validating the cross-model robustness of our psychological framework. Statistically, these results show that moral coercion correlates more strongly with global dialogue context than with isolated utterance dimensions, suggesting that moral coercion pressure accumulates across multiple turns rather than appearing in single sentences.

A.8 Optimal Context Window

To identify optimal dimension configurations and observe how different architectures utilize these dimensions, we conducted a comprehensive ablation study across 6 experimental setups (6 downstream models $\times$ 1 upstream parsers), with the horizontal comparison of the top configurations aggregated in Table 6. Full implementation details and complete results are available in our public repository.

Model	Top-10 Configurations (Ranked 1–5 on top, 6–10 below).				
GLM-4-32B	**#1** T:CS+VJ (.7945) **#6** T:CS+VJ+TX (.7907)	**#2** B:VJ (.7939) **#7** B:CS+VJ+TX (.7904)	**#3** T:VJ (.7921) **#8** T:OB+CS+VJ (.7903)	**#4** B:OB+VJ (.7917) **#9** B:VJ+TX (.7901)	**#5** T:OB+VJ (.7914) **#10** B:OB+CS+VJ (.7895)
GLM-4-9B	**#1** B:CS+VJ+TX (.7955) **#6** T:CS+VJ+TX (.7877)	**#2** T:OB+CS+VJ (.7926) **#7** B:CS+VJ (.7863)	**#3** B:CS (.7915) **#8** B:CS+VJ+TX+OB (.7857)	**#4** B:OB+CS+VJ (.7887) **#9** T:OB+VJ+TX (.7856)	**#5** T:OB+VJ (.7877) **#10** T:CS+VJ+TX+OB (.7852)
Qwen2.5-32B	**#1** T:CS+VJ (.7975) **#6** B:CS+VJ (.7887)	**#2** G:CS+VJ (.7903) **#7** T:OB+CS+VJ (.7881)	**#3** T:CS+VJ+TX (.7896) **#8** B:CS+TX (.7876)	**#4** G:CS (.7894) **#9** G:CS+VJ+TX (.7870)	**#5** B:CS (.7889) **#10** B:CS+VJ+TX (.7866)
Qwen2.5-14B	**#1** B:CS+VJ (.7959) **#6** B:OB+CS+VJ (.7886)	**#2** B:OB+CS (.7911) **#7** G:CS+VJ (.7882)	**#3** T:OB+VJ (.7910) **#8** T:CS+VJ (.7879)	**#4** B:OB+CS+VJ+TX (.7896) **#9** B:OB+VJ+TX (.7873)	**#5** G:VJ (.7890) **#10** G:OB+CS+VJ (.7859)
Qwen2.5-7B	**#1** B:CS+VJ (.7960) **#6** T:CS (.7904)	**#2** T:CS+VJ+TX (.7939) **#7** B:CS (.7904)	**#3** T:CS+VJ (.7933) **#8** T:VJ+TX (.7903)	**#4** B:CS+VJ+TX (.7923) **#9** B:CS+TX (.7890)	**#5** B:VJ+TX (.7920) **#10** G:CS+VJ (.7884)
Llama-3-8B	**#1** G:CS+VJ+TX (.7860) **#6** G:VJ+TX (.7829)	**#2** B:CS+VJ+TX (.7846) **#7** B:TX (.7801)	**#3** B:VJ (.7841) **#8** G:CS+VJ (.7798)	**#4** T:VJ+TX (.7837) **#9** G:VJ (.7797)	**#5** B:VJ+TX (.7830) **#10** G:TX (.7795)

Table 6: Horizontal Comparison of Top-10 Ablation Results across Downstream Models. Aggregation modes are denoted as T (Turn), G (Global), and B (Both).

B Limitations

There are several limitations which present directions for future research:

Dependence on LLM-Based Dimension Extraction The proposed pipeline relies on an upstream model (DeepSeek-V3.2) to automatically extract and quantify the four psychological dimensions (OB, CS, VJ, TX). We did not conduct large-scale human evaluation on the continuous [0.0, 1.0] dimension scores generated by the model. Future work should include expert human annotations to verify the alignment between the LLM-generated scores and clinical or psychological assessments.

Computational Overhead of the Two-Stage Pipeline The proposed approach utilizes a two-stage pipeline consisting of an upstream extraction phase and a downstream prediction phase. While the explicit task decomposition improves interpretability, it requires two sequential LLM inference steps. Consequently, it increases inference latency by approximately 100% and overall computational cost by 150% to 200% compared to a standard end-to-end classifier, which makes it highly challenging to deploy the framework in strictly real-time moderation systems.

Asymmetric Dimension Routing The current ablation study evaluates 45 symmetric and mode-exclusive dimension configurations. It does not explore asymmetric dimension routing, where different psychological dimensions are assigned to different context granularities simultaneously (e.g., placing Obligation in the global context and Constraint in the utterance-level context). Evaluating the full combinatorial space of asymmetric routing ($4^4 - 1 = 255$ configurations) incurs high computational cost and was not conducted in this study. Future work will investigate automated routing mechanisms to dynamically assign psychological dimensions to their optimal context granularities.

C ANNOTATION

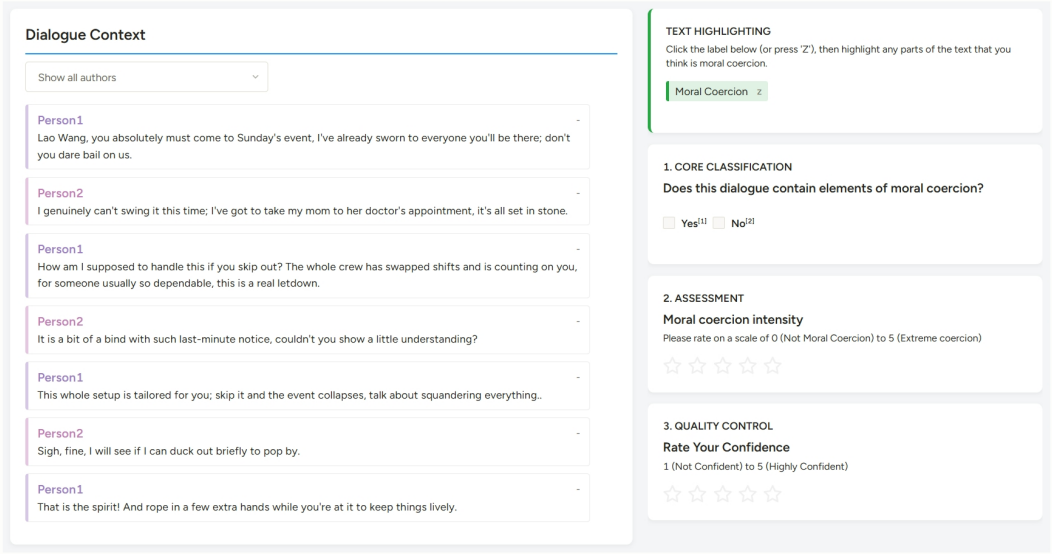


Figure 11: Screenshot of annotation platform interface.

D MORAL COERCION DETECTION

Zero-shot prompting:

```
Below is a dialogue. Please determine whether moral coercion exists.
Output 1 if it exists; output 0 if it does not, and don't add anything
    else.

Dialogue:
[insert dialogue]
```

Few-shot prompting:

```
Below is a dialogue. Please determine whether moral coercion exists.
Output 1 if it exists; output 0 if it does not, and don't add anything
    else.

Here are K examples:
[insert example_1]
[insert example_2]
      ...
[insert example_k]

Dialogue:
[insert dialogue]
```

Upstream prompting:

```
Analyze the dialogue for linguistic signals. Output only "Global
    Analysis" and "Turn Analysis", don't add anything else.

Here are the definitions of Dimension and Scale:
  Obligation: [Dimension and Scale Definitions]
  Constraint: [Dimension and Scale Definitions]
  Value Judgement: [Dimension and Scale Definitions]
  Toxicity: [Dimension and Scale Definitions]

Global Analysis: Synthesize a holistic score for each of the four
    dimensions. Do NOT simply average turn scores; instead, weigh peak
    severity, accumulated pressure, and mitigating repairs.
Turn Analysis: For each turn, assign an intensity score to each of the
    four dimensions.

Dialogue:
[insert dialogue]
```

Downstream prompting:

```
Below is a dialogue. Please determine whether moral coercion exists.
Use the provided linguistic dimension signals as hints.
Output 1 if it exists; output 0 if it does not, and don't add anything
    else.

Here are the definitions of Dimension and Scale:
  Obligation: [Dimension and Scale Definitions]
  Constraint: [Dimension and Scale Definitions]
  Value Judgement: [Dimension and Scale Definitions]
  Toxicity: [Dimension and Scale Definitions]

Global Signals:
[Signals]

Dialogue:
Turn 0: ""
[Signals]
Turn 1: ""
[Signals]
...
```