Patterns of Persuasion Through the Lens of Theory of Mind: Value Alignment Analysis in Online Deliberation

Baktash Ansari[1*] **Mouly Dewan**[2] **Vibhor Agarwal**[3] **Afra Mashhadi**[1]

[1] Computing and Software Systems, University of Washington, Bothell, WA, USA [2]Information School, University of Washington, Seattle, WA, USA [3]Nokia Bell Labs, Cambridge, United Kingdom

Abstract

Understanding what makes an argument persuasive is central to computational social science. Yet, most approaches rely on surface-level linguistic features or uninterpretable neural classifiers. We propose a framework that examines persuasion through the lens of *Theory of Mind* (ToM): the cognitive capacity to model others' beliefs, desires, intentions, emotions, knowledge, and perspectives. Using an LLM to extract structured ToM value profiles from 19,340 post-comment pairs on the `r/ChangeMyView` subreddit, we compute fine-grained alignment features between posts and their responses. A human annotation study provides initial face-validity evidence for the extraction. Our analysis surfaces four interpretable patterns that distinguish persuasive from non-persuasive responses at the population level: a cognitive–affective split, in which persuasive comments align more with a post's cognitive dimensions (beliefs, desires) while highly-engaged but unpersuasive comments rely more on affective dimensions (emotions); a "cover and reframe" tendency, addressing the author's emotional and factual concerns while introducing novel intentional framing; a directional cost of lexical echoing, where exact overlap in knowledge tokens carries the strongest negative coefficient in a logistic model; and an internal-consistency effect, where persuasive comments show somewhat more unified ToM profiles. We frame the contribution as a measurement framework and hypothesis-generating empirical study that lays groundwork for ToM-informed evaluation of LLM social reasoning, not as a persuasion classifier.

Introduction

Persuasion, the act of changing someone's beliefs or attitudes through argument, is central to deliberative discourse. Online platforms such as Reddit's `r/ChangeMyView` (CMV) provide a unique setting where persuasion outcomes are explicitly labeled by the original poster, offering a rare window into how minds change in public discourse.

Understanding these dynamics has become newly urgent in the age of large language models. A growing body of work has moved beyond treating Theory of Mind (ToM), the ability to attribute mental states such as beliefs, desires, intentions, and emotions to others, as a standalone evaluation benchmark, and instead embeds mental state reasoning into model inputs, prompt structures, and multi-agent architectures to generate more adaptive and socially coherent responses (Wilf et al. 2024). Yet this line of research faces a fundamental bottleneck: we lack the evaluation metrics needed to assess whether LLMs are genuinely reasoning about persuasion in human-like ways. Cognitive science has long established that ToM is a prerequisite for effective persuasion (Slaughter, Peterson, and Moore 2013; Bartsch and London 2000). Children with stronger ToM abilities select arguments tailored to the persuadee's mental state (Bartsch and London 2000), and affective ToM (emotion attribution) directly predicts persuasion skill (Zonca et al. 2022). Yet LLMs have shown consistent limitations in zero-shot ToM reasoning requiring nuanced integration of human mental states (Ma et al. 2023; Zhou et al. 2023; Street 2024). Building such evaluation metrics requires first understanding how persuasion actually operates in human online discourse.

Prior computational work has relied on surface-level linguistic features such as hedging, lexical diversity, and sentiment (Tan et al. 2016), or end-to-end classifiers that trade interpretability for predictive power (Habernal and Gurevych 2016), neither of which grounds persuasion in the mental state reasoning that cognitive science identifies as its core mechanism. This paper addresses that gap by connecting comment-level ToM signals to observable persuasion outcomes in online deliberation, providing the empirical foundation from which richer, cognitively grounded evaluation metrics for LLM social reasoning can be derived.

Concretely, we address the following research question: **do persuasive comments align differently with the ToM values of the original post compared to non-persuasive ones, and if so, along which cognitive dimensions?** We identify four interpretable patterns that distinguish persuasive from non-persuasive responses at the population level: a cognitive–affective split, a cover-and-reframe tendency, a directional cost of lexical echoing, and an internal coherence effect. These patterns are aggregate tendencies, not individual-level predictors: within-post discrimination is near chance and our logistic model reaches AUC ≈ 0.52. We therefore frame the contribution as a measurement framework and hypothesis-generating empirical study, connecting individual persuasive exchanges to the broader challenge

*Corresponding author: baktash@uw.edu.

of evaluating social reasoning in online communities and in LLMs. The contributions of this paper are as follows:

- We propose a pipeline for extracting structured ToM value profiles from argumentative text using an LLM across 6 cognitive dimensions, with a human face-validity study on 30 samples (2 annotators; 88% validity rate, 82% inter-annotator agreement).
- We analyze 19,340 post-comment pairs across 3,040 posts and identify four patterns that separate persuasive from non-persuasive responses: a cognitive-affective split, a cover-and-reframe strategy, a cost of lexical echoing, and an internal coherence effect.
- We show that ToM alignment features provide clear, theory-grounded signals that go beyond surface-level linguistic features, with internal coherence showing the strongest effect across all features, offering a foundation for future evaluation tools for LLM social reasoning.

Related Work

We review related work, organized into two research dimensions: Theory of Mind and its importance in NLP, and persuasion in online discourse.

Theory of Mind in NLP

Theory of Mind (ToM) refers to the cognitive capacity to attribute mental states, such as beliefs, desires, intentions, and emotions, to oneself and others, and to recognize that these states may differ across individuals (Leslie, Friedman, and German 2004). First formalized through the classic "false belief" paradigm (Wimmer and Perner 1983), ToM is fundamental to human social cognition, enabling people to interpret communicative intent, anticipate behavior, and engage in cooperative and persuasive interactions.

In NLP, ToM is increasingly recognized as a critical capability for modeling language use in social contexts. Understanding text often requires going beyond surface-level semantics to infer implicit beliefs, goals, and perspectives of authors and their readers in online communities. Recent works have therefore examined whether large language models (LLMs) exhibit ToM-like reasoning abilities. Van Duijn et al. (2023) compared 11 LLMs with children aged 7–10 on complex ToM tasks, showing that instruction-tuned GPT models often outperform children, while simpler models struggle. Sap et al. (2022) established benchmarks for social intelligence in LLMs, finding significant limitations in mental-state inference. Furthermore, Wu et al. (2023) found that ToM understanding declines as task complexity increases, limiting models' performance and Shapira et al. (2024) showed that LLM ToM abilities are brittle under distribution shifts or adversarial conditions. Wilf et al. (2024) proposed SimToM, a perspective-taking framework that improves LLM ToM capabilities through simulation-based prompting. BigToM (Gandhi et al. 2023) and ToMi (Le, Boureau, and Nickel 2019) focus on perspective-taking and false-belief reasoning, while OpenToM (Xu et al. 2024) and FANToM (Kim et al. 2023) assess narrative and conversational mental-state tracking. While much of the prior work focuses on evaluating whether LLMs possess ToM capabilities, an emerging direction treats LLMs as tools for extracting structured representations of mental states from text. In this view, LLMs can be used to infer latent variables such as an author's beliefs, intentions, or emotional stance (Bojić et al. 2025; Quan et al. 2026), effectively operationalizing ToM as a set of measurable features. In this work, we use LLMs to derive structured ToM representations from human-written discourse, and then leverage them to analyze patterns of persuasion.

Persuasion in Online Discourse

Persuasion in online discourse is inherently a social-cognitive process that depends not only on the content of an argument but also on how well it engages with the audience's underlying mental states. Early work on `r/ChangeMyView` demonstrated that both linguistic features (e.g., tone, hedging, lexical diversity) and interaction dynamics (e.g., timing, interplay) are predictive of persuasive success (Tan et al. 2016). However, later studies emphasize that persuasion is fundamentally constrained by the recipient's prior beliefs and attitudes, which often outweigh surface linguistic features (Durmus and Cardie 2018). Additionally, Hidey et al. (2017) demonstrated that combinations of classical persuasive appeals (ethos, logos, pathos) contribute to successful arguments.

Despite these insights, most prior computational work focuses on observable features of persuasive messages rather than explicitly modeling the mental states of participants and the alignment between them. We argue that modeling *mental model alignment* between participants is central to understanding online persuasion. Successful arguments are those that effectively bridge the gap between the author's and the audience's beliefs and intentions. Building on this perspective, our work leverages ToM representations extracted from text to quantify and analyze how such alignment evolves between participants in online conversations, providing a more cognitively grounded account of persuasive success.

Data

To study persuasion at scale, we focus on `r/ChangeMyView`. In CMV, users can post about a variety of topics and request arguments that help them change their view. The original poster (OP) then awards a "delta" (Δ) to a comment that changed their view, creating a verifiable ground-truth signal.

Collection and Cleaning

We collected all `r/ChangeMyView` posts and comments spanning February 2013 through January 2026 via the Arctic Shift's download tool[1]. As the official Reddit API does not provide historical Reddit data, we use this open-source tool which maintains historical Reddit data in dumps.

Delta detection. We identified posts where the OP awarded a delta by scanning all comments for delta markers (`!delta`, Δ, or HTML equivalents) in replies authored

[1]https://arctic-shift.photon-reddit.com/download-tool

by the OP. This yielded 40,451 posts with at least one delta-awarded comment.

Cleaning. We removed posts where the body, title, or author had been deleted or removed by Reddit, and posts where the persuasive comment itself had been deleted. This produced a clean dataset of **38,064 posts**, each with one identified persuasive comment. The distribution of posts by year is shown in Figure 1.

Multi-delta posts. Because our analysis is structured around a single persuasive vs. non-persuasive contrast per post, we keep one persuasive comment per post and treat the rest as follows: for each post, we select the first non-deleted delta-awarded comment encountered when streaming the Arctic Shift comment dump (a stable but order-dependent choice, not strictly the earliest or highest-scored delta), and we exclude all other delta-awarded comments and their corresponding OP delta replies from the negative-comment pool so that they do not contaminate the hard or easy negative groups. If the first encountered persuasive comment was deleted, we move to the next; posts where every delta-awarded comment was deleted are dropped.

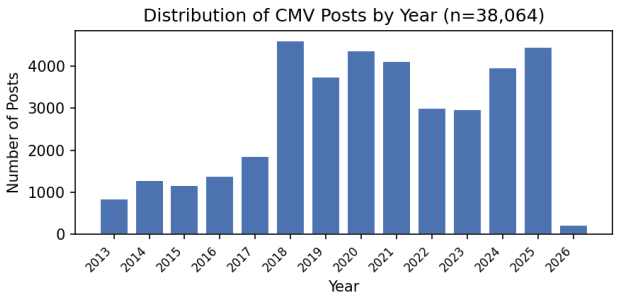

Figure 1: Distribution of 38,064 `r/ChangeMyView` posts by year (February 2013 to January 2026).

Temporal filtering and Comment selection. To ensure that non-persuasive comments were not influenced by the winning argument, we retained only comments posted before the persuasive comment's timestamp. From each post's remaining comments, we excluded bot comments (AutoModerator, DeltaBot) and categorized the rest according to their votes by users. It is worth noting that the most upvoted comment is not necessarily the persuasive (delta) comment, as our data confirms that 81% of the highly upvoted comments (hard negatives) are not the persuasive comments. For every post, Reddit provides a score that represents the net popularity of a post, calculated as total upvotes minus total downvotes. Based on these Reddit score quartiles, we construct three comparison groups:

- **Persuasive** ($n = 3{,}040$): Successful arguments. These are comments where the original author explicitly stated their mind was changed (awarded a "delta"). To maintain balance, the dataset includes exactly one successful comment per post.
- **Hard negative** ($n = 8{,}150$): High-effort but unsuccessful arguments. These are highly upvoted comments (in the top 25% of Reddit scores) that made strong, substantive points but ultimately failed to persuade the author. A single post can have multiple comments in this category (average 2.68 per post). Notably, 81% of hard negative comments outscore the persuasive comment in the same post, and in 91% of posts at least one hard negative has a higher score than the delta-awarded comment, confirming that community approval and persuasion are distinct phenomena.
- **Easy negative** ($n = 8{,}150$): Low-effort unsuccessful arguments. These are poorly received comments (in the bottom 25% of Reddit scores), serving as a baseline for weak engagement.

Comments with average scores (the middle 50%) were excluded to create a stark contrast between strong (hard) and weak (easy) failures. This three-tier design enables distinguishing between two distinct signals: *engagement* (persuasive/hard negative vs. easy negative) and *persuasion* (persuasive vs. hard negative).

Analysis Subset

We extracted ToM values for a random subset of 4,000 posts. Of these, 3,040 had complete ToM profiles for the persuasive comment and at least one hard and one easy negative, yielding **19,340 post-comment pairs** (3,040 persuasive + 8,150 hard negative + 8,150 easy negative). Table 1 summarizes the analysis subset.

Table 1: Analysis subset statistics (19,340 pairs from 3,040 posts).

		Word Count		Reddit Score	
Type	**Count**	**Mean**	**Med.**	**Mean**	**Med.**
Posts	3,040	345	262	–	–
Persuasive	3,040	191	137	18.8	3
Hard neg	8,150	129	89	40.1	12
Easy neg	8,150	90	56	−0.2	1

Persuasive comments are substantially longer than both negative types (median 137 vs. 89 and 56 words), consistent with prior findings that argument length correlates with persuasion (Tan et al. 2016). Hard negatives have the highest median Reddit score (12 vs. 3 for persuasive), reinforcing that community approval does not equate to persuasive success. The full word count distributions are shown in Figure 2.

ToM Value Extraction

Extracting human values from text has roots in Schwartz's value theory (Schwartz 2012). Kiesel et al. (2022) introduced a 54-value taxonomy for argument mining grounded in Schwartz theory. Ziems et al. (2024) systematically evaluated LLMs as zero-shot annotators for computational social science tasks, finding reasonable agreement with human annotators. For similarity computation, we draw on sentence-transformer embeddings (Reimers and Gurevych

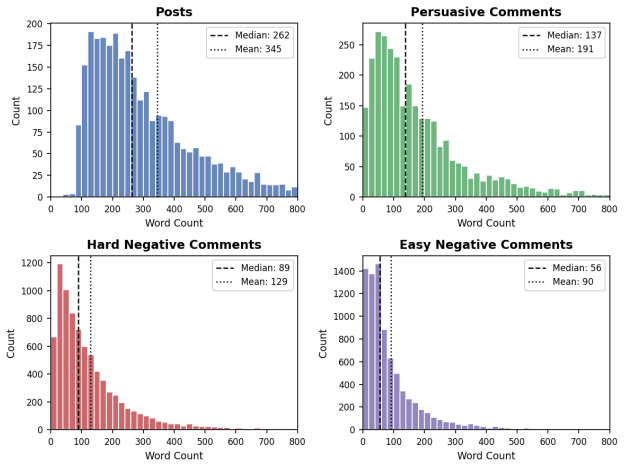

Figure 2: Word count distributions for posts and the three comment types in the analysis subset (3,040 posts, 19,340 pairs).

2019) and token-level cosine similarity methods related to BERTScore (Zhang et al. 2020).

Our approach operationalizes ToM as a structured set of keyword-level value tokens attributed to an author's mental states by an LLM; these are a computational proxy for human mental state reasoning, not a direct measure of ToM capacity. We use Qwen3-30B-A3B (Qwen Team 2025) to extract ToM values from each post and comment across 6 dimensions (Table 2). Inference was performed using vLLM on a server with 3 NVIDIA RTX A5000 GPUs (24 GB VRAM each) with temperature $= 0.6$, top-p $= 0.95$, and max tokens $= 1{,}024$; no random seed was fixed (generation is non-deterministic). To avoid truncation, we use a two-run strategy: the first run extracts beliefs, desires, and intentions; the second extracts emotions, knowledge, and perspective-taking. The results are merged into a single profile per text. When the model returns `none` for a category, the token list for that category is treated as empty, yielding a zero vector for cosine similarity and a novelty score of 1.0. The complete extraction prompts are provided in Appendix A.

Table 2: ToM categories and definitions.

Category	Definition
Beliefs	What the author thinks is true/-false
Desires	What the author wants or values
Intentions	What the author plans or advocates
Emotions	Emotional states or appeals
Knowledge	What the author knows or assumes
Perspective-taking	How other viewpoints are considered

For each category, the model produces a comma-separated list of value keywords (e.g., `Empathy, Accountability, Fairness`) and a brief content description. Extraction produces 46,740 unique value tokens across the full dataset.

To support reproducibility and future research, we release the curated dataset of 19,340 post-comment pairs and the accompanying analysis code in a public repository.[2]

Human Annotation Study

To assess the face validity of LLM-extracted ToM values, we conduct a human annotation study. The study is intended as initial validation rather than a definitive reliability assessment; we discuss its scope and caveats below.

Setup

We randomly sample 30 post-comment pairs from the persuasive subset. Each sample includes the original post with its ToM analysis, the persuasive comment with its ToM analysis, and the OP's delta reply for context. Annotators access samples through a custom annotation interface built for this study.[3]

Annotation Task

For each sample, annotators answer 6 binary (Yes/No) questions:

- **Q1–Q6 (ToM Validity)**: For each of the 6 ToM categories, do the extracted values for the persuasive comment make sense given the comment text?

Annotators are instructed to judge whether the extracted values are *supported by* the comment text, not whether the extraction is exhaustive. When a category has no clear content in the comment, annotators mark "Yes" if the extraction is appropriately minimal. Full annotation guidelines are provided in Appendix B.

Annotators

Two annotators participated in the study. The annotators were graduate students of computer science and information science. Annotation was conducted on a voluntary basis and no financial compensation was provided. The annotators were provided the code book and annotation guideline described above.

Results

We report two complementary measures in Table 3. Let $N = 30$ be the number of annotated samples and let $y_{i,r}^{(c)} \in \{0, 1\}$ denote annotator r's Yes/No judgment on category c for sample i, where 1 indicates the extraction was judged valid.

Validity rate. This measures how often the LLM extraction was considered supported by the comment text, pooled across both annotators:

$$\text{Validity}^{(c)} = \frac{1}{2N} \sum_{i=1}^{N} \sum_{r \in \{1,2\}} y_{i,r}^{(c)}.$$

[2] https://github.com/baktash81/persuasion-tom

[3] https://annotation.baktashans.com/

The aggregate validity rate is the mean of $\text{Validity}^{(c)}$ over the six categories.

Percent agreement. This measures raw inter-annotator agreement, i.e., the fraction of items on which the two annotators gave the same Y/N label:

$$\text{Agree}^{(c)} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[y_{i,1}^{(c)} = y_{i,2}^{(c)} \right].$$

The aggregate percent agreement is the mean of $\text{Agree}^{(c)}$ over the six categories.

Table 3: Human validation of ToM extraction on 30 persuasive comments with 2 annotators. Validity = pooled fraction of Y judgments; % Agree. = raw inter-annotator agreement.

Category	Validity (%)	% Agree.
Q1: Beliefs	90.0	86.7
Q2: Desires	91.7	83.3
Q3: Intentions	93.3	86.7
Q4: Emotions	86.7	73.3
Q5: Knowledge	95.0	90.0
Q6: Perspective-taking	73.3	73.3
Aggregate	88.3	82.2

Across the 30 samples, annotators judged the extracted values as supported in 88.3% of all binary decisions and agreed with each other on 82.2% of items. Perspective-taking shows the lowest values on both measures, reflecting the more interpretive nature of that category.

We emphasize three scope limitations of this study: (i) the sample is small (30 comments), (ii) it covers only persuasive comments and therefore measures face validity on the easy-positive case rather than across the full distribution including hard and easy negatives, and (iii) the metric is whether extracted keywords are *supported by* the text, not whether they are exhaustive or canonical. We treat the results as initial face-validity evidence for the extraction pipeline, not as a definitive construct-validity claim. A larger, prevalence-balanced annotation that includes negatives is an important direction for future work and would also enable informative chance-corrected reliability statistics.

Method

Our analysis pipeline has three stages: token embedding, per-category aggregation, and feature computation.

Token Embedding

Each category's value string is split on the commas and lowercased into individual tokens. All unique tokens are embedded using `all-MiniLM-L6-v2` (Reimers and Gurevych 2019), producing 384-dimensional, L2-normalized vectors. This enables measuring semantic similarity between value concepts even when different words are used (e.g., "empathy" $\leftrightarrow$ "compassion").

Per-Category Mean Pooling

For a given post-comment pair and ToM category c, let $\{t_1, \ldots, t_n\}$ be the post's tokens with embeddings $\{\vec{e}_1, \ldots, \vec{e}_n\}$. The category-level representation is as follows:

$$\vec{v}^{(c)} = \frac{\bar{\vec{e}}}{\|\bar{\vec{e}}\|_2}, \quad \bar{\vec{e}} = \frac{1}{n}\sum_{i=1}^{n} \vec{e}_i \tag{1}$$

This creates a semantic centroid for each category, normalized to unit length so that cosine similarity equals the dot product.

Feature Computation

For each pair and each of the 6 categories, we compute four alignment features between the post's token set $P^{(c)}$ and the comment's token set $C^{(c)}$:

Cosine Similarity. Semantic alignment of value concepts: $\text{sim}^{(c)} = \vec{v}^{(c)}_{\text{post}} \cdot \vec{v}^{(c)}_{\text{comment}}$

Jaccard Overlap. Exact token overlap: $J^{(c)} = |P^{(c)} \cap C^{(c)}|/|P^{(c)} \cup C^{(c)}|$

Value Novelty. Fraction of comment values absent from the post: $\text{nov}^{(c)} = |C^{(c)} \setminus P^{(c)}|/|C^{(c)}|$

Value Coverage. Fraction of post values addressed by the comment: $\text{cov}^{(c)} = |P^{(c)} \cap C^{(c)}|/|P^{(c)}|$

This yields 24 alignment features (6 categories $\times$ 4 metrics).

Statistical testing. We use two-sided Mann-Whitney U tests, which require no distributional assumptions, and report Cohen's d as the effect-size measure. Because multiple comments may originate from the same post, we additionally report Benjamini-Hochberg (BH) FDR-adjusted q-values within each test family in Tables 9 and 10, and we cross-validate the logistic regression model with `GroupKFold` grouped by post id (Appendix D) so that no post contributes comments to both training and validation folds.

Internal Consistency

Inspired by ROSCOE's step-to-step consistency metric (Golovneva et al. 2023), which measures whether adjacent reasoning steps are semantically coherent, we introduce an analogous measure for ToM profiles: *internal consistency*. Instead of checking coherence between reasoning steps, we check whether the 6 ToM categories within a single comment form a unified mental model.

For a given comment, let $\vec{v}^{(c)}$ be the mean-pooled value vector for category c, computed via the same embedding and pooling procedure described in Eq. 1. While the alignment features compare post vs. comment vectors for the *same* category, internal consistency compares the comment's own vectors *across different* categories. Concretely, we compute pairwise cosine similarity between all $\binom{6}{2} = 15$ category pairs:

$$\text{cons}(c_i, c_j) = \vec{v}^{(c_i)} \cdot \vec{v}^{(c_j)} \tag{2}$$

We then aggregate into four summary scores:

- **Overall consistency**: mean of all 15 pairwise scores.

- **Cognitive consistency**: mean over the 6 pairs among cognitive categories (beliefs, desires, intentions, knowledge).
- **Affective consistency**: cosine between emotions and perspective-taking.
- **Cross-domain consistency**: mean of the 8 pairs crossing cognitive and affective categories.

We compute the same overall consistency for the post ($\text{cons}_{\text{post}}$) and define **consistency difference** as:

$$\Delta\text{cons} = \text{cons}_{\text{comment}} - \text{cons}_{\text{post}} \quad (3)$$

Consistency difference measures whether the comment presents a more or less coherent ToM profile than the post it responds to. This yields 15 pairwise + 6 aggregate = 21 additional features.

Results

The Cognitive-Affective Split

Table 4 shows per-category cosine similarity between post and comment ToM value embeddings.

Table 4: Mean cosine similarity by ToM category and comment type, with p-values from Mann-Whitney U tests. P–H = persuasive vs. hard negative; P–E = persuasive vs. easy negative. * $p < .05$, *** $p < .001$.

				p-value	
Category	**Easy**	**Hard**	**Pers.**	**P–H**	**P–E**
Beliefs	.489	.490	**.493**	.556	.082
Desires	.501	.507	**.510**	.231	**<.001*** **
Intentions	.473	.477	**.483**	**.030***	**<.001*** **
Emotions	.575	**.583**	.580	.267	.099
Knowledge	.448	.458	**.459**	.697	**<.001*** **
Perspective	.475	**.485**	.481	.201	.094

A directional pattern emerges in the group means: persuasive comments are slightly higher on all four cognitive dimensions (beliefs, desires, intentions, knowledge), while hard negatives are slightly higher on both affective dimensions (emotions, perspective-taking). The differences are small and most are not significant under MWU+BH-FDR (Appendix Table 9); we read this as a directional tendency at the population level, suggestive of successful persuasion engaging more with *what the OP thinks and wants* relative to over-investing in emotional mirroring, rather than as a confirmed mechanism.

The intentions category shows the largest persuasive advantage ($\Delta = +0.006$) and is the only feature reaching raw statistical significance in the persuasive vs. hard negative comparison ($p = 0.030$, Mann-Whitney U; Cohen's $d = 0.046$); the effect is small and does not survive BH-FDR over the 24-test family ($q = .72$). When comparing persuasive vs. easy negative, three cognitive categories reach high significance (desires, intentions, knowledge; all $p < 0.001$, all $q < .005$ after BH-FDR), while neither affective dimension does ($p > 0.05$), giving the cognitive–affective split most of its statistical support in the engaged-vs-disengaged contrast rather than in the harder persuasive-vs-hard-negative one. Full statistical test results are provided in Tables 9 and 10 in the Appendix.

Implications. In online communities, upvotes and emotional resonance are often mistaken for persuasive quality. Our results suggest that platforms aiming to promote genuine opinion change should surface responses that engage with what the other person believes and wants, rather than those that simply match their emotional tone. This has implications for how discussion platforms rank and recommend replies in deliberative spaces.

Cover and Reframe

Value coverage analysis (Table 5) reveals that persuasive comments have the highest coverage on emotions, desires, and intentions, while hard negatives lead only on perspective-taking coverage.

Table 5: Mean value coverage (fraction of post values addressed) by category and type. P–H = persuasive vs. hard negative; P–E = persuasive vs. easy negative. * $p < .05$, ** $p < .01$.

				p-value	
Category	**Easy**	**Hard**	**Pers.**	**P–H**	**P–E**
Beliefs	.041	.040	**.041**	.713	.714
Desires	.044	.044	**.047**	.391	.368
Intentions	.041	.041	**.045**	.280	.348
Emotions	.126	.135	**.140**	.384	**.004****
Knowledge	.029	.034	**.034**	.795	**.037***
Perspective	.065	**.072**	.070	.472	.200

Table 6 shows value novelty, the fraction of comment value tokens absent from the post. Novelty is uniformly high (85–97%) across all types and categories, meaning comments introduce new value vocabulary rather than reusing the post's terms.

Table 6: Mean value novelty (fraction of comment values absent from post) by category and type. P–H = persuasive vs. hard negative; P–E = persuasive vs. easy negative. * $p < .05$.

				p-value	
Category	**Easy**	**Hard**	**Pers.**	**P–H**	**P–E**
Beliefs	.951	.953	.952	.693	.803
Desires	.950	.951	.949	.432	.502
Intentions	.953	.955	.953	.345	.495
Emotions	.851	.844	.847	.875	.150
Knowledge	.968	.963	**.964**	.866	**.048***
Perspective	.927	.921	.927	.250	.703

Combined, these tables reveal a "cover and reframe" pattern. The statistical significance concentrates in the persuasive vs. easy negative comparison: persuasive comments show significantly higher *coverage* on emotions ($p = 0.004$) and knowledge ($p = 0.037$), and significantly lower knowledge *novelty* ($p = 0.048$), indicating they address more of

the post's emotional and factual concerns using partially overlapping vocabulary. In contrast, no coverage or novelty feature reaches significance in the persuasive vs. hard negative comparison (all $p > 0.25$). This indicates that the "cover and reframe" pattern primarily distinguishes *engaged* from *disengaged* responses: both persuasive and hard negative comments acknowledge the post's concerns at similar rates, while easy negatives fail to engage with the post's mental state. The pattern captures a necessary condition for persuasion (engaging with the OP's concerns) rather than a sufficient one (actually changing their mind). At the same time, novelty remains above 85% across all categories and comment types, confirming that even when comments address the post's values, they do so using predominantly new framing rather than echoing the post's vocabulary.

Implications. This pattern suggests that effective participation in online discussions requires both acknowledging the other person's concerns and offering a new perspective. Simply ignoring what the original poster said (low coverage) tends to produce low-quality engagement. Community platforms could use this insight to guide users toward more constructive replies, particularly in spaces designed for deliberation or conflict resolution.

Echoing Hurts

Jaccard overlap is uniformly low (2–9%), confirming that posts and comments rarely share exact value tokens. To quantify the directional association of each feature with persuasion, we examine logistic regression coefficients from a model trained on all 24 alignment features (Table 7). The model is used here as an analytical tool to identify directional associations, not as a predictor: its cross-validated AUC of 0.525 on the persuasive vs. hard negative task confirms it has limited predictive power, consistent with the non-significant Mann-Whitney results. Full model details are provided in Tables 12 and 13 in the Appendix.

Table 7: Selected logistic regression coefficients (standardized) chosen for interpretability. Positive = associated with persuasion. Full ranked coefficients in Table 13.

Feature	Coef.	Interpretation
emotions_cov	+0.43	Cover post's emotions
knowledge_cov	+0.33	Cover post's knowledge
emotions_nov	+0.27	New emotional framing
intentions_cov	+0.23	Address post's goals
intentions_nov	+0.22	Introduce new goals
knowledge_jacc	−0.39	**Echo knowledge hurts**
beliefs_cov	−0.13	Agree w/ beliefs hurts
emotions_jacc	−0.11	Echo emotions hurts

The strongest negative coefficient is `knowledge_jaccard` (-0.39): in this directional analysis, exact overlap in knowledge-related value tokens is the largest single indicator pointing *against* persuasion. Conversely, the largest positive coefficients (`emotions_coverage` $+0.43$, `knowledge_coverage` $+0.33$) suggest that *semantically* engaging with the post's concerns is associated with persuasion, even though the comment does not echo the same vocabulary. We emphasize that these are coefficient-level associations from a low-AUC model and should be read as hypothesis-generating signal rather than as evidence of a causal mechanism.

Implications. At the population level, exact lexical mirroring of a post's knowledge tokens is associated with reduced odds of persuasion. If this directional signal replicates in causal designs, it would suggest that community writing tools and AI assistants should encourage engaging with the same topic in fresh language rather than echoing the original wording.

Engagement vs. Persuasion Signal

To validate that ToM features capture meaningful signal, we compare persuasive vs. easy negative comments. Multiple features significantly separate the groups ($p < 0.001$): knowledge similarity ($d = 0.082$), intentions similarity ($d = 0.074$), and desires similarity ($d = 0.071$). This confirms that ToM value alignment captures real *engagement* signal. However, these features do not significantly separate persuasive from hard negative comments (all $p > 0.05$ except intentions, $p = 0.03$), indicating that the ToM signal primarily reflects topical relevance rather than persuasive effectiveness per se. Full results are in Appendix C.

Paired Analysis

To ensure the observed differences are not simply driven by topic variation (for instance, an emotional post naturally drawing emotional replies regardless of persuasiveness), we compare persuasive and hard negative comments *within the same post*. For posts that contain both types, we ask: does the persuasive comment consistently score higher on ToM alignment than its hard negative counterpart?

The answer is no. The fraction of cases where the persuasive comment scores higher stays near chance across all six categories (range: 0.475–0.502), and the standard deviation of pairwise differences is 50–80$\times$ larger than the mean difference. This means ToM alignment cannot reliably pick the winning comment in a given post.

The significant group-level differences identified in earlier analyses (e.g., the cognitive-affective split) are therefore best understood as *population-level tendencies*: consistent patterns across thousands of pairs that describe how successful arguments are generally structured, rather than a rule that holds for every individual case.

Coherent Mental Models

While other patterns examine *inter-text* alignment (post vs. comment), we now examine *intra-text* coherence: do the 6 ToM categories within a comment form a unified mental model? This is analogous to ROSCOE's step-to-step consistency metric, which checks whether reasoning steps cohere with each other.

Table 8 reports all 15 pairwise cosine similarities between ToM category vectors within a comment, plus aggregate

scores, with p-values for both persuasive vs. hard negative (P–H) and persuasive vs. easy negative (P–E).

Table 8: Internal ToM consistency: all 15 pairwise cosine similarities and aggregate scores. Easy/Hard/Pers = group means. P–H: persuasive vs. hard negative; P–E: persuasive vs. easy negative. * $p < .05$, ** $p < .01$, *** $p < .001$.

	Mean Consistency			p-value	
Feature	**Easy**	**Hard**	**Pers.**	**P–H**	**P–E**
bel↔des	.550	.555	.551	ns	ns
bel↔int	.492	.500	**.500**	ns	**
bel↔emo	.469	.470	**.477**	ns	**
bel↔kno	.541	**.556**	.554	ns	***
bel↔per	.453	.462	**.464**	ns	***
des↔int	.501	.506	**.510**	ns	***
des↔emo	.448	.454	**.461**	*	***
des↔kno	.463	**.468**	.466	ns	ns
des↔per	.444	**.460**	**.460**	ns	***
int↔emo	.430	.437	**.444**	**	***
int↔kno	.445	.449	**.453**	ns	***
int↔per	.453	.463	**.468**	*	***
emo↔kno	**.408**	.401	.401	ns	*
emo↔per	.465	**.470**	.469	ns	*
kno↔per	**.429**	.428	.427	ns	ns
Overall cons.	.466	.472	**.474**	ns	***
Cognitive cons.	.499	**.506**	**.506**	ns	***
Affective cons.	.465	**.470**	.469	ns	*
Cross-domain	.442	.447	**.450**	ns	***
Post cons.	**.481**	**.481**	.478	**	**
Δcons	−.015	−.009	**−.004**	**	***

Persuasive vs. Hard Negative. Comparing persuasive and hard negative comments, overall consistency does not reach raw significance ($p = 0.232$); a few specific cross-domain pairs do reach raw $p < .05$: int↔emo ($p = 0.006$), des↔emo ($p = 0.012$), and int↔per ($p = 0.049$), and the consistency difference reaches $p = 0.003$. These point in the direction that persuasive comments are slightly more coherent *relative to their post*. Effect sizes are small and the signal weakens further once family-wise corrections are applied; we therefore treat these as directional tendencies in the harder P-vs-H comparison.

Persuasive vs. Easy Negative. The signal strengthens substantially. Overall consistency ($p < 10^{-7}$), cross-domain consistency ($p < 10^{-6}$), and consistency difference ($p < 10^{-9}$) are all highly significant and survive BH-FDR correction. The consistency difference is the largest effect observed across all ToM features in this study. The strongest pairwise signals involve *intentions* or *desires* crossing into the affective domain: des↔per, int↔emo, and int↔per all reach $p < .001$. Persuasive comments integrate what they want to achieve with how they frame it emotionally, while low-engagement comments show more scattered ToM profiles. We note that this comparison primarily distinguishes engaged from disengaged commenters; the persuasion-specific (P-vs-H) signal on the same metric is weaker.

The fact that Δcons is stronger than raw comment consistency suggests the effect is not simply that persuasive comments are inherently more coherent, but that they are more coherent *relative to the complexity of the post they address*.

Implications. Persuasive online arguments tend to present a clear and unified point of view, where the commenter's goals, reasoning, and emotional tone all point in the same direction. This suggests that tools designed to support online deliberation could use internal coherence as a signal of argument quality, helping users and moderators identify well-structured contributions in discussions where many responses are scattered or contradictory.

Discussion

Theoretical Implications

Elaboration Likelihood Model. The cognitive-affective split aligns with the distinction between central-route (argument quality) and peripheral-route (emotional appeal) persuasion (Petty and Cacioppo 1986). Persuasive comments align more on cognitive dimensions, consistent with central-route processing in the high-engagement `r/ChangeMyView` context where users are motivated to evaluate arguments carefully.

Theory of Mind and Persuasion. Our "cover and reframe" pattern mirrors developmental findings that children with better ToM select persuasive arguments tailored to the target's mental state (Bartsch and London 2000). Effective persuaders in `r/ChangeMyView` similarly demonstrate awareness of the OP's concerns (high coverage on emotions and knowledge) while offering a novel perspective (high novelty on intentions).

The Cost of Echoing. The finding that exact lexical overlap in knowledge values *hurts* persuasion complements Durmus and Cardie (2018)'s insight that prior beliefs mediate persuasion. Simply restating what the OP already knows does not challenge their view; it reinforces it.

Coherent Mental Models. The internal consistency insight, that persuasive comments have more coherent ToM profiles, particularly across cognitive-affective boundaries, resonates with research on argument quality. Effective arguments are not simply collections of good points; they present an integrated worldview where goals, emotions, beliefs, and evidence reinforce each other. The strongest pairwise signals (intentions↔emotions, desires↔perspective) suggest that persuasive commenters successfully connect *what they advocate* with *how they frame it emotionally*, producing a unified persuasive narrative.

Toward a ToM-Informed Evaluation Suite

This work sets the foundation for future evaluation framework research. Inspired by ROSCOE (Golovneva et al. 2023), which uses unsupervised semantic similarity metrics to assess step-by-step reasoning quality focusing on evaluating *logical* reasoning coherence, we envision using the foundations proposed in this paper to evaluate *social* reasoning. Future work will include measuring alignment between the

mental-state profiles of a post and its response. Such measures reframe persuasion analysis as a structured comparison of cognitive representations, rather than a bag-of-words or end-to-end prediction task.

1. **Cognitive Coverage**: Does the response address the target's stated beliefs, desires, intentions, and knowledge?
2. **Affective Acknowledgment**: Does it engage with the target's emotional state?
3. **Intentional Novelty**: Does it introduce new goals or framings the target had not considered?
4. **Echo Avoidance**: Does it avoid restating the target's existing knowledge?
5. **Internal Coherence**: Does the response present a unified mental model where cognitive and affective dimensions reinforce each other?

The patterns identified in this paper provide empirically grounded criteria for such a suite. While our current features have limited predictive power for individual comments, they offer *interpretable dimensions* along which persuasive and non-persuasive responses systematically differ. Concretely, these dimensions can be applied to evaluate LLM-generated responses: given a prompt post, one can extract ToM profiles from both the post and an LLM-generated reply, then score the reply along the five criteria above to assess whether it exhibits the same ToM-coherent structure observed in human persuaders—providing a principled, theory-grounded complement to surface fluency metrics.

Limitations and Future Work

Conversation chain context. Our analysis treats each post-comment pair in isolation and does not account for the broader thread. In practice, a comment's persuasive success may partly depend on prior exchanges in the conversation. Future work should incorporate full thread context to better attribute what drives opinion change.

Computational cost of ToM extraction. Extracting ToM profiles with Qwen3-30B-A3B is resource-intensive. Each post requires inference over the post itself and all associated comments (at least 8 inference passes per post), which limited our analysis to a 4,000-post subset of the full 38,064-post dataset. Future work should explore lighter extraction models or distillation approaches to scale the framework.

ToM extraction quality and validation scope. Values are extracted by a single LLM, introducing potential noise. The high vocabulary sparsity (46,740 unique tokens across 19,340 pairs) may inflate novelty and deflate Jaccard scores. Our annotation study provides only initial face-validity evidence: it covers 30 persuasive comments with 2 annotators and measures whether extracted keywords are supported by the text, not whether they are exhaustive or canonical. Hard and easy negatives are not included in the validation sample, so we cannot rule out that extraction quality differs across comment types. A larger-scale human evaluation, ideally with more annotators, balanced across all three comment groups, and using a richer category-level rubric, is an important direction for future work.

Value keyword granularity. Representing ToM values as keyword lists is a coarse approximation. Persuasion also depends on reasoning structure, evidence quality, and rhetorical strategy, none of which are captured at this level. Future work could enrich the representation with structured argument graphs or richer semantic features.

Delta as persuasion proxy. Delta awards are an imperfect signal: some OPs award them liberally, others rarely, and the first sufficient argument may receive a delta regardless of overall quality. Future work could complement delta-based labels with other persuasion signals such as OP reply sentiment or view-change self-reports.

Population vs. individual signal. The patterns we identify are population-level tendencies and do not reliably predict persuasion for individual post-comment pairs. Closing this gap between aggregate patterns and individual-level prediction is an important direction for future work.

Conclusion

We presented a ToM-based framework for analyzing persuasion patterns in online deliberation. By extracting structured value profiles across 6 cognitive dimensions and computing fine-grained alignment and consistency features, we identified four interpretable population-level patterns that distinguish persuasive from non-persuasive responses: a cognitive–affective split favoring cognitive alignment in persuasive comments, a "cover and reframe" tendency combining emotional/knowledge coverage with intentional novelty, a directional cost of echoing where exact lexical overlap in knowledge tokens carries the strongest negative coefficient in a low-AUC logistic model, and an internal-consistency effect where persuasive comments show somewhat more unified ToM profiles. The signal is most consistent in the persuasive vs. low-engagement comparison; for the harder persuasive vs. high-engagement (hard-negative) comparison, group-level differences are small and within-post discrimination is near chance, so we treat these patterns as aggregate tendencies rather than individual-level predictors. A human annotation study (30 samples, 2 annotators; 88% validity rate, 82% inter-annotator agreement) provides initial face-validity evidence for the underlying ToM extraction.

These findings are not intended as a standalone persuasion predictor. Instead, they provide interpretable, theory-grounded dimensions that complement existing linguistic approaches and lay the groundwork for ToM-informed evaluation suites for assessing the quality of LLM-generated social reasoning. Future work will integrate these ToM features with argument structure analysis and develop a comprehensive evaluation framework for persuasive LLM outputs.

Acknowledgments

This work was supported in part by a Royalty Research Fund award from the University of Washington Office of Research. The authors declare no competing interests. Generative AI tools were used for proofreading portions of the manuscript text and for formatting table content; all scientific content, analysis, and conclusions are the authors' own.

References

Bartsch, K.; and London, K. 2000. Children's Use of Mental State Information in Selecting Persuasive Arguments. *Developmental Psychology*, 36(3): 352–365.

Bojić, L.; Zagovora, O.; Zelenkauskaite, A.; Vuković, V.; Čabarkapa, M.; Veseljević Jerković, S.; and Jovančević, A. 2025. Comparing large Language models and human annotators in latent content analysis of sentiment, political leaning, emotional intensity and sarcasm. *Scientific reports*, 15(1): 11477.

Durmus, E.; and Cardie, C. 2018. Exploring the Role of Prior Beliefs for Argument Persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 1035–1045. Association for Computational Linguistics.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gandhi, K.; Fränken, J.-P.; Gerstenberg, T.; and Goodman, N. 2023. Understanding social reasoning in language models with language models. *Advances in Neural Information Processing Systems*, 36: 13518–13529.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Golovneva, O.; Chen, M.; Poff, S.; Corredor, M.; Zettlemoyer, L.; Fazel-Zarandi, M.; and Celikyilmaz, A. 2023. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Habernal, I.; and Gurevych, I. 2016. What Makes a Convincing Argument? Empirical Analysis and Detecting Attributes of Convincingness in Web Argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1214–1223. Association for Computational Linguistics.

Hidey, C.; Musi, E.; Hwang, A.; Muresan, S.; and McKeown, K. 2017. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining (ArgMining@EMNLP)*, 11–21. Association for Computational Linguistics.

Kiesel, J.; Alshomary, M.; Handke, N.; Cai, X.; Wachsmuth, H.; and Stein, B. 2022. Identifying the Human Values behind Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4459–4471. Association for Computational Linguistics.

Kim, H.; Sclar, M.; Zhou, X.; Bras, R.; Kim, G.; Choi, Y.; and Sap, M. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14397–14413.

Le, M.; Boureau, Y.-L.; and Nickel, M. 2019. Revisiting the evaluation of theory of mind through question answering. 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)*, 5872–5877.

Leslie, A. M.; Friedman, O.; and German, T. P. 2004. Core mechanisms in 'theory of mind'. *TRENDS in Cognitive Sciences*, 8(12).

Ma, Z.; Sansom, J.; Peng, R.; and Chai, J. 2023. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1011–1031. Singapore: Association for Computational Linguistics.

Petty, R. E.; and Cacioppo, J. T. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer-Verlag.

Quan, X.; Xiong, J.; Valentino, M.; and Freitas, A. 2026. Inferring Latent Intentions: Attributional Natural Language Inference in LLM Agents. *arXiv preprint arXiv:2601.08742*.

Qwen Team. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*, 3982–3992. Association for Computational Linguistics.

Sap, M.; LeBras, R.; Fried, D.; and Choi, Y. 2022. Neural Theory-of-Mind? On the Limits of Social Intelligence in Large Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3762–3780. Association for Computational Linguistics.

Schwartz, S. H. 2012. An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2(1).

Shapira, N.; Levy, M.; Alavi, S. H.; Zhou, X.; Choi, Y.; Goldberg, Y.; Sap, M.; and Shwartz, V. 2024. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2257–2273. Association for Computational Linguistics.

Slaughter, V.; Peterson, C. C.; and Moore, C. 2013. I Can Talk You into It: Theory of Mind and Persuasion Behavior in Young Children. *Developmental Psychology*, 49(2): 227–231.

Street, W. 2024. LLM Theory of Mind and Alignment: Opportunities and Risks. *arXiv preprint arXiv:2405.08154*.

Tan, C.; Niculae, V.; Danescu-Niculescu-Mizil, C.; and Lee, L. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-Faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, 613–624. International World Wide Web Conferences Steering Committee.

Van Duijn, M.; Van Dijk, B.; Kouwenhoven, T.; De Valk, W.; Spruit, M.; and van der Putten, P. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th conference on computational natural language learning (CoNLL)*, 389–402.

Wilf, A.; Lee, S.; Liang, P. P.; and Morency, L.-P. 2024. Think Twice: Perspective-Taking Improves Large Language Models' Theory-of-Mind Capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 8292–8308. Association for Computational Linguistics.

Wimmer, H.; and Perner, J. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1): 103–128.

Wu, Y.; He, Y.; Jia, Y.; Mihalcea, R.; Chen, Y.; and Deng, N. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 10691–10706.

Xu, H.; Zhao, R.; Zhu, L.; Du, J.; and He, Y. 2024. OpenToM: A comprehensive benchmark for evaluating theory-of-mind reasoning capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8593–8623.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhou, P.; Zhu, A.; Hu, J.; Pujara, J.; Ren, X.; Callison-Burch, C.; Choi, Y.; and Ammanabrolu, P. 2023. I Cast Detect Thoughts: Learning to Converse and Guide with Intents and Theory-of-Mind in Dungeons and Dragons. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11136–11155. Toronto, Canada: Association for Computational Linguistics.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1): 237–291.

Zonca, J.; Vignaud, P.; Franck, N.; and Holler, J. 2022. Persuasion Ability in Children: Relations to Cognitive and Affective Theory of Mind. *Frontiers in Psychology*, 13: 966102.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. Our study analyzes publicly available Reddit discussions to understand persuasion patterns at a population level. It does not involve profiling individuals, perpetuate unfair stereotypes, or disrespect any group.

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes. The abstract and introduction accurately describe our framework, dataset, the four identified patterns, and their scope as population-level tendencies rather than individual-level predictors.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. Method section describes the feature computation pipeline in detail. We use non-parametric Mann-Whitney U tests appropriate for non-normal distributions, and report Cohen's d effect sizes alongside p-values.

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. We discuss potential artifacts in the Limitations and Future Work section, including the self-selected nature of r/ChangeMyView participants and the use of delta awards as an imperfect persuasion proxy.

   (e) Did you describe the limitations of your work? Yes. See the Limitations and Future Work subsection in the Discussion.

   (f) Did you discuss any potential negative societal impacts of your work? No, because our findings describe population-level structural patterns and do not provide a functional tool for manipulation.

   (g) Did you discuss any potential misuse of your work? No. Our framework has limited individual-level predictive power (AUC $\approx$ 0.52), which substantially reduces potential for misuse as a manipulation tool.

   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. We use only publicly available Reddit data accessed through the Arctic Shift API in compliance with Reddit's data access policies. The data does not include private communications or sensitive personal information. Full hyperparameters and analysis details are provided in the Appendix to support reproducibility.

   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.

2. Additionally, if your study involves hypotheses testing...

   (a) Did you clearly state the assumptions underlying all theoretical results? Yes. We use non-parametric Mann-Whitney U tests, which require no distributional assumptions. The experimental design, including the three-tier comparison and quartile-based comment selection, is fully described in the Data section.

   (b) Have you provided justifications for all theoretical results? Yes. All reported patterns are supported by Mann-Whitney U tests with p-values and Cohen's d

effect sizes. Full statistical results are provided in Tables 9 and 10 in the Appendix.

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Yes. The Discussion section relates our findings to the Elaboration Likelihood Model and prior work on computational persuasion, discussing both supporting and complementary theoretical frameworks.

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes. Our paired within-post analysis specifically tests whether observed group-level differences are artifacts of topic variation rather than genuine persuasion signals.

(e) Did you address potential biases or limitations in your theoretical framework? Yes. The Limitations and Future Work section addresses biases including delta award noise, vocabulary sparsity, keyword-level granularity, and the gap between population-level patterns and individual-level prediction.

(f) Have you related your theoretical results to the existing literature in social science? Yes. The Related Work and Discussion sections connect our findings to prior work in computational persuasion, Theory of Mind research in NLP, and the Elaboration Likelihood Model from social psychology.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes. Each result subsection includes an Implications paragraph, and the Toward a ToM-Informed Evaluation Suite subsection outlines concrete directions for future research and platform design.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA. This paper does not include theoretical proofs.

(b) Did you include complete proofs of all theoretical results? NA. This paper does not include theoretical proofs.

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. Analysis code, processed data, and replication instructions are available at https://github.com/baktash81/persuasion-tom.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. Full hyperparameters for the logistic regression model are provided in Table 12 in the Appendix. Feature construction is described in Method section.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes. Cross-validation AUC is reported as mean $\pm$ standard deviation over 5 folds (AUC = 0.525 ± 0.006).

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes. TOM extraction was performed using vLLM on a server with 3 NVIDIA RTX A5000 GPUs (24 GB VRAM each). These details are reported in the Limitations and Future Work section.

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes. We justify the use of non-parametric tests, the three-tier comparison design, and explicitly note that the logistic regression is used as an analytical tool to identify directional associations rather than as a classifier, consistent with its near-chance AUC.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes. The Paired Analysis section discusses that TOM alignment cannot reliably identify the winning comment in an individual post (near-chance accuracy), and the Limitations section notes that our patterns are population-level tendencies rather than individual predictors.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes. We cite Arctic Shift for data collection, Qwen3-30B-A3B for TOM extraction, all-MiniLM-L6-v2 for embeddings, and all other tools and datasets used.

(b) Did you mention the license of the assets? Yes. The ToM Value Extraction section links the release and states licenses (MIT for code, CC BY-NC 4.0 for redistributed data). We cite Arctic Shift, Qwen3, and sentence-transformers.

(c) Did you include any new assets in the supplemental material or as a URL? Yes. The dataset and code are at https://github.com/baktash81/persuasion-tom. The annotation interface is at https://annotation.baktashans.com/.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes. We use publicly posted Reddit content, which users submit under Reddit's terms of service that permit academic research use. No private data was accessed.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? No. Our dataset consists of publicly posted Reddit content analyzed at an aggregate level. We do not retain or analyze individual user identities, and individual posts are not reproduced in the paper.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes. The release includes `FAIR.md` with a concise FAIR-oriented summary.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al.

(2021))? Yes. The release includes `DATASHEET.md` following the Gebru et al. datasheet template.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes. Full annotation guidelines are provided in the Appendix. The annotation interface including detailed instructions was built for this study and it is available on https://annotation.baktashans.com
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? No. The annotation task involves reading publicly available Reddit posts and answering binary questions, which poses minimal risk to participants. IRB approval was not sought as the task does not involve sensitive data or vulnerable populations.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes. We state in the Human Annotation Study section that the annotation was conducted on a voluntary basis and no financial compensation was provided.
   (d) Did you discuss how data is stored, shared, and deidentified? No. Annotation responses are stored on a private server. No personally identifiable information about annotators is collected beyond voluntary participation.

A ToM Extraction Prompts

Due to output length constraints, we use a two-run extraction strategy. The prompts below show the two-part version used in production.

Post Analysis: Part 1 (Beliefs, Desires, Intentions)

```
You are analyzing a post from
r/ChangeMyView (CMV) through Theory
of Mind (TOM).
<post>
Title: {title}
Body: {body}
</post>
Analyze ONLY these three TOM categories
(keep output brief):
1. BELIEFS: What the author thinks is
  true/false. Values (1-4 words each,
  comma-separated) and Content (brief).
2. DESIRES: What the author wants/values.
  Values and Content.
3. INTENTIONS: What the author plans or
  wants others to do. Values and Content.
Output Format:
<analysis>
2-3 relatively short sentences of
analysis and reasoning.
</analysis>
<beliefs>
Values: Value1, Value2 (or "none")
Content: Brief description
</beliefs>
<desires>
Values: Value1, Value2 (or "none")
Content: Brief description
</desires>
<intentions>
Values: Value1, Value2 (or "none")
Content: Brief description
</intentions>
IMPORTANT: Output only these four tags.
Be concise.
```

Post Analysis: Part 2 (Emotions, Knowledge, Perspective)

```
You are analyzing a post from
r/ChangeMyView (CMV) through Theory
of Mind (TOM).
<post>
Title: {title}
Body: {body}
</post>
Analyze ONLY these three TOM categories
(keep output brief):
4. EMOTIONS: What the author feels or
  attributes to others. Values and
  Content.
5. KNOWLEDGE: What the author knows/
  doesn't know. Values and Content.
6. PERSPECTIVE-TAKING: How the author
  considers other viewpoints. Values
  and Content.
Output Format:
<emotions>
Values: Value1, Value2 (or "none")
Content: Brief description
</emotions>
<knowledge>
Values: Value1, Value2 (or "none")
Content: Brief description
</knowledge>
<perspective_taking>
Values: Value1, Value2 (or "none")
Content: Brief description
</perspective_taking>
IMPORTANT: Output only these three tags.
Be concise.
```

Comment Analysis

Comment prompts follow the same two-part structure but include the post title as context:

```
<context>
Original Post Title: {title}
</context>
<comment>
{body}
</comment>
```

The rest of the prompt is identical to the post version.

B Annotation Guidelines

The annotation study uses 30 samples from the persuasive subset, accessed via a custom web interface. Each sample shows the original post (with ToM analysis), the persuasive comment (with ToM analysis), and the OP's delta reply.

Questions

Q1 *Beliefs*: Do the listed belief-values match what the commenter seems to believe or assert?

Q2 *Desires*: Do the desire-values match what the commenter wants or prioritizes?

Q3 *Intentions*: Do the intention-values match what the commenter is trying to do or achieve?

Q4 *Emotions*: Do the emotion-values match the emotional tone or appeals in the comment?

Q5 *Knowledge*: Do the knowledge-values match what the commenter claims to know or how they use evidence?

Q6 *Perspective-taking*: Do the perspective-taking values match how the commenter considers other viewpoints?

All questions are binary (Yes/No). An optional free-text note field is provided per sample. Annotators can flag bad instances. Detailed instructions with criteria for Yes/No judgments are provided in the annotation interface.

C Full Statistical Results

Columns for all tables below: **P/H/E** = group means, d = Cohen's d effect size, p = p-value from Mann-Whitney U test, **Sig.** = significance level (* $p < .05$, ** $p < .01$, *** $p < .001$, ns = not significant).

Persuasive vs Hard Negative

Table 9 reports Mann-Whitney U tests for all 24 alignment features.

Persuasive vs Easy Negative

Table 10 reports statistically significant alignment features only.

Paired Analysis (Within-Post)

For posts with both persuasive and hard negative comments:

Standard deviations ($\sim$0.16) are 50–80$\times$ larger than mean differences ($\sim$0.002), confirming that noise vastly overwhelms any per-post signal.

D Logistic Regression Model Details

We trained a binary logistic regression classifier to distinguish persuasive from hard negative comments using all 24 ToM alignment features. Features were z-scored before fitting and performance evaluated via 5-fold stratified cross-validation (AUC = 0.525 ± 0.006). To rule out optimistic estimates from comments of the same post leaking across folds, we additionally evaluated with 5-fold `GroupKFold` grouped by post id (AUC = 0.518 ± 0.004); the two estimates agree to within one standard deviation, confirming that clustering does not materially change the conclusion that the model has limited individual-level predictive power. Table 12 lists the hyperparameters and Table 13 reports the full standardized coefficients.

Table 9: Mann-Whitney U: persuasive vs hard negative (all 24 alignment features). q = Benjamini-Hochberg FDR-adjusted p-value across the 24-test family. Sig. refers to raw p (* $p < .05$).

Feature	P	H	d	p	q	Sig.
beliefs_sim	.493	.490	.017	.556	.875	ns
desires_sim	.510	.507	.024	.231	.797	ns
intentions_sim	.483	.477	.046	.030	.724	*
emotions_sim	.580	.583	−.020	.267	.797	ns
knowledge_sim	.459	.458	.010	.697	.875	ns
perspective_sim	.481	.485	−.027	.201	.797	ns
beliefs_jacc	.028	.027	.008	.708	.875	ns
desires_jacc	.030	.029	.015	.407	.797	ns
intentions_jacc	.029	.027	.020	.314	.797	ns
emotions_jacc	.095	.095	.004	.764	.875	ns
knowledge_jacc	.023	.023	−.007	.840	.875	ns
perspective_jacc	.047	.050	−.023	.377	.797	ns
beliefs_nov	.952	.953	−.009	.693	.875	ns
desires_nov	.949	.951	−.012	.432	.797	ns
intentions_nov	.953	.955	−.011	.345	.797	ns
emotions_nov	.847	.844	.014	.875	.875	ns
knowledge_nov	.964	.963	.006	.866	.875	ns
perspective_nov	.927	.921	.032	.250	.797	ns
beliefs_cov	.041	.040	.006	.713	.875	ns
desires_cov	.047	.044	.020	.391	.797	ns
intentions_cov	.045	.041	.029	.280	.797	ns
emotions_cov	.140	.135	.024	.384	.797	ns
knowledge_cov	.034	.034	.003	.795	.875	ns
perspective_cov	.070	.072	−.016	.472	.809	ns

Table 10: Mann-Whitney U: persuasive vs easy negative (alignment features with raw $p < .05$). q = Benjamini-Hochberg FDR-adjusted p across the full 24-test family. Sig. refers to raw p.

Feature	P	E	d	p	q	Sig.
desires_sim	.510	.501	.071	$<.001$	.003	***
intentions_sim	.483	.473	.074	$<.001$	.001	***
knowledge_sim	.459	.448	.082	$<.001$	.001	***
emotions_jacc	.095	.090	.033	.037	.145	*
knowledge_jacc	.023	.020	.031	.043	.145	*
emotions_cov	.140	.126	.063	.004	.024	**
knowledge_cov	.034	.029	.042	.037	.145	*
knowledge_nov	.964	.968	$-.031$	.048	.145	*

Table 11: Paired cosine similarity: persuasive minus hard negative (same post).

Category	Mean Δ	Std	P > H
Beliefs	-0.002	0.173	0.475
Desires	-0.000	0.163	0.502
Intentions	-0.001	0.167	0.486
Emotions	-0.003	0.153	0.495
Knowledge	-0.003	0.176	0.490
Perspective	-0.002	0.161	0.481

Table 12: Logistic regression hyperparameters.

Parameter	Value
Regularization	L2, $C = 1.0$
Class weight	Balanced
Max iterations	1,000
CV folds	5 (stratified; also `GroupKFold` by post id)
Input features	24 (z-scored)
Training samples	10,924

Table 13: Full logistic regression coefficients (standardized), sorted by absolute value. Positive = associated with persuasion.

Feature	Coef.
emotions_coverage	$+0.434$
knowledge_jaccard	-0.394
knowledge_coverage	$+0.326$
emotions_novelty	$+0.269$
intentions_coverage	$+0.232$
intentions_novelty	$+0.218$
perspective_coverage	$+0.216$
perspective_novelty	$+0.152$
beliefs_coverage	-0.130
desires_coverage	$+0.118$
emotions_jaccard	-0.111
beliefs_novelty	-0.085
perspective_jaccard	-0.084
desires_jaccard	-0.079
emotions_sim	-0.070
intentions_sim	$+0.048$
knowledge_novelty	-0.044
beliefs_jaccard	$+0.039$
desires_novelty	$+0.033$
intentions_jaccard	-0.029
knowledge_sim	$+0.013$
perspective_sim	-0.013
beliefs_sim	$+0.009$
desires_sim	$+0.007$