Persuasion Index: A Theory-Guided Framework for Persuasion Analysis

Anonymous ACL submission

Abstract

Identifying persuasive rhetorical cues is critical across domains, from detecting information manipulation and improving AI safety to advancing public health communication. We propose **Persuasion Index** (PI), a taxonomy of 15 dimensions grounded in persuasion theories from psychology and communication, and one transparent implementation using 55 subfeatures built from lexicons and rule-based detectors. The taxonomy is modular: individual detectors can be replaced while preserving the theoretical structure. By evaluating PI on four public datasets varying in domain, style, and outcome measures, we show that PI provides a shared feature space for interpreting rhetorical patterns associated with persuasion-related outcomes. Linear models show that PI features carry meaningful predictive signal while remaining computationally lightweight. Dimension-level analyses reveal recurring associations between PI dimensions and persuasion outcomes across datasets, while also highlighting topic- and stance-specific variation. We release PI as an open-source package and web interface for principled and auditable analysis of human and AI-mediated communication.[1]

1 Introduction

Persuasion is the process whereby a message changes an audience's beliefs, attitudes, or behavior (O'Keefe, 2015). Understanding persuasion has become increasingly urgent as AI systems gain influence over public discourse at scale (Salvi et al., 2025; Hackenburg and Margetts, 2024; Bozdag et al., 2026). LLMs have demonstrated measurable attitude change in large-scale political persuasion experiments (Hackenburg et al., 2025b); AI-generated content accelerates the spread of health misinformation (Augenstein et al., 2024); and conversational agents increasingly operate as persuaders in both human-facing and multi-agent systems (Breum et al., 2024; Matz et al., 2024; Dash et al., 2025; Cheng and You, 2026). In these settings, detecting persuasion is not enough; auditing it requires identifying the underlying rhetorical cues deployed, as logical, emotional, and credibility-based appeals call for distinct responses.

Figure 1: Two arguments against plastic pollution. Message A (**Logos**-heavy) relies on factual grounding through statistics and institutional citation; Message B (**Pathos**-heavy) relies on emotional engagement, urgency, and threat framing.

Prior NLP work on persuasion has largely focused on predicting persuasive outcomes rather than explaining the rhetorical mechanisms behind them. Persuasive outcome varies widely across datasets, ranging from pairwise convincingness judgments (Habernal and Gurevych, 2016) to self-reported opinion change (Tan et al., 2016) and experimentally measured pre–post attitude shifts (Durmus et al., 2024). Models ranging from bag-of-words classifiers to fine-tuned transformers and LLM-based judges can achieve strong predictive performance, but their outputs are typically binary labels or scalar scores (Yang et al., 2019; Bassi et al., 2024). For example, Figure 1 shows two arguments reaching the same conclusion but deploying entirely different rhetorical cues, a distinction that scalar persuasiveness predictions absorb but cannot expose. While post-hoc attribution methods can explain which tokens or features influenced a

[1] Our code, data, website, and package will all be publicly-available upon publication.

particular model, they do not provide a stable vocabulary for comparing persuasive language across datasets, domains, or communicative settings. This creates a need for an interpretable framework that represents persuasive language directly in terms of theory-grounded rhetorical dimensions, independent of any particular prediction model or dataset-specific label scheme.

We address this gap by proposing the **Persuasion Index** (PI) with two distinct components. The core component is a taxonomy of 15 dimensions grounded in well-established theories from psychology, communication, and rhetoric, mapped onto the Aristotelian triad of **Logos**, **Ethos**, and **Pathos**. The second component is an implementation of this taxonomy through 55 granular sub-features, implemented via theory-seeded lexicons, LLM-based expansion, validated computational linguistics resources, and lightweight rule-based detectors. PI is modular by design: we provide a complete implementation, but individual sub-feature measurements are interchangeable and extensible without disrupting the taxonomy. Unlike opaque neural representations, PI extracts features directly from text, ensuring every score traces back to specific lexical or structural patterns. We make the following contributions:

1. The Persuasion Index (PI), a theory-guided taxonomy of persuasive rhetoric
2. An implementation via interpretable lexicon and rule-based features
3. Empirical predictive validation across four heterogeneous persuasion datasets
4. Analysis of PI dimensions and persuasive outcomes, revealing both regularities and variation across datasets, topics, and stances
5. An open-source package and public web interface for real-time argument analysis (Appendix J), enabling researchers to apply and extend the framework on their own data.

The PI can broadly support interdisciplinary persuasion research, with use cases including content analysis or experimental stimuli validation in social science research and AI safety audits of persuasive content in human-LLM or multi-agent interactions.

2 Background

Research in psychology and communication has long treated persuasion as multidimensional rather than a single scalar property. Dual-process theories such as the Elaboration Likelihood Model and the Heuristic-Systematic Model distinguish between *central/systematic processing*, characterized by rigorous evaluation of evidence and argument quality, and *peripheral/heuristic processing*, characterized by reliance on cues such as source credibility and emotional tone (Petty and Cacioppo, 1986; Chaiken, 1980). PI builds on this tradition by including rhetorical dimensions corresponding to both routes; Section 3 details the theoretical anchors for each PI dimension.

Prior NLP work has largely operationalized persuasion through task- and domain-specific taxonomies, each emphasizing a narrow subset of persuasive mechanisms. Research on altruistic requests, crowdfunding and charitable giving foregrounds peripheral cues with an emphasis on social and relational strategies (Althoff et al., 2014; Mitra and Gilbert, 2014; Wang et al., 2019; Yang et al., 2019; Chen and Yang, 2021). Conversely, argument mining research primarily targets logic-oriented argument structure and quality (Wachsmuth et al., 2017; Ghosh et al., 2016; Carlile et al., 2018; Gretz et al., 2020; Toledo et al., 2019; Joshi et al., 2023; Romberg et al., 2025). Propaganda detection research defines and labels fine-grained persuasive techniques in news and political texts (Da San Martino et al., 2019; Dimitrov et al., 2021; Piskorski et al., 2023; Modzelewski et al., 2025; Sajwani et al., 2024).

A second source of fragmentation is that datasets define persuasive success in heterogeneous ways, including pairwise convincingness judgments (Habernal and Gurevych, 2016; Gleize et al., 2019; Toledo et al., 2019; Gretz et al., 2020), scalar ratings of argument quality (Ghosh et al., 2016; Carlile et al., 2018; Joshi et al., 2023). A complementary paradigm measures pre–post attitude change, using explicit markers of opinion shifts or survey responses (Tan et al., 2016; Luu et al., 2019; Monti et al., 2022; Durmus et al., 2024). Thus, existing resources provide strong local accounts of persuasive language, but their domain-specific taxonomies and heterogeneous outcome labels make it difficult to compare how persuasive language across domains or determine if observed patterns generalize to other corpora.

Like many other tasks, NLP research often emphasizes predictive performance over explanatory power, modeling has trended toward higher-capacity, lower-interpretability architectures. Early work combines hand-crafted linguistic features, argument structure, and interaction dynamics with

linear or tree-based models (Stab and Gurevych, 2014a; Somasundaran et al., 2007; Danescu-Niculescu-Mizil et al., 2013; Althoff et al., 2014; Mitra and Gilbert, 2014; Ghosh et al., 2016; Wang et al., 2017). Later studies adopt neural architectures for predicting persuasiveness, argument quality, and opinion change (Hidey and McKeown, 2018; Yang et al., 2019; Wang et al., 2019; Toledo et al., 2019; Gretz et al., 2020; Chen and Yang, 2021; Joshi et al., 2023). More recently, LLMs act as persuaders in controlled experiments, where their messages are evaluated via human attitude shifts (Breum et al., 2024; Matz et al., 2024; Bai et al., 2025; Hackenburg and Margetts, 2024; Hackenburg et al., 2025a; Salvi et al., 2025; Cheng and You, 2026; Costello et al., 2024; Dash et al., 2025; Sharma et al., 2024), and as judges of persuasive strength and strategy use (Pauli et al., 2025; Breum et al., 2024; Salvi et al., 2025; Durmus et al., 2024; Modzelewski et al., 2025). As LLMs increasingly generate, personalize, and evaluate persuasive messages, it is necessary for NLP research to not just predict persuasion, but provide interpretable measures of underlying rhetorical strategies.

Recent work has begun to address this need by grounding predictions in theory. Sudharsan and Pacheco (2025) show that incorporating argumentative components and semantic types improve cross-domain generalization. Hoang et al. (2025) combine psychological theory with LLM-generated feature ratings in interpretable classifiers. PI shares this motivation but differs in scope and operationalization. Rather than focusing primarily on argumentative structure, PI represents persuasive language across logical, affective, and credibility appeals. While the PI framework is modular, the implementation we provide uses transparent lexicon and rule-based features. These choices make PI a shared, auditable representation for comparing persuasive language across contexts and describing linguistic strategies beyond outcome prediction.

3 Theoretical-Grounded Framework

The PI maps its 15 dimensions onto Aristotelian Triad of **Logos**, **Ethos**, and **Pathos** (Aristotle, 2007), a meta-framework for organizing the predictors of persuasion, also used in argument quality and rhetorical analysis (Carlile et al., 2018; Higgins and Walker, 2012). Under **Logos**, PI targets the central route of cognitive elaboration, including EVIDENCE, SPECIFICITY, LOGIC AND COHESION, ARGUMENTATION, and OPPONENT'S VIEW. Under **Ethos**, PI captures heuristic credibility signals, including AUTHORITY AND CREDIBILITY, POLITENESS, COMMITMENT, and STYLE. Under **Pathos**, PI contains affective and social compliance triggers, including SENTIMENT, IMPACT, ENGAGEMENT, RECIPROCITY, SCARCITY AND URGENCY, and PROPAGANDA. Figure 2 shows the 15 dimensions and 55 sub-features, with theoretical anchors and prior computational linguistic operationalizations mapped in Table A1.

3.1 Logos: The Appeal to Reason

Logos captures arguments' internal consistency and factual grounding, targeting the central route of cognitive processing where receivers carefully evaluate claim validity (Petty and Cacioppo, 1986).

EVIDENCE: verifiable facts, statistics, and attributed information used to ground claims. Sub-features target cognitive beliefs by anchoring claims in verifiable facts and authoritative origins (Zebregs et al., 2015). The inclusion of named entities is supported by Exemplification Theory: specific examples enhance perceived issue severity and message recall (Bigsby et al., 2019).

LOGIC AND COHESION: the argument's reasoning is explicitly structured and cognitively accessible through connective and referential language. High text cohesion facilitated by logical connectives reduces cognitive load, allowing readers to focus on content rather than processing structure (Kaakinen et al., 2011). This structural logic serves as a cognitive facilitator ensuring messages can be processed via systematic routes (Chaiken, 1980).

ARGUMENTATION: structural explicitness of claims and the quantity and sophistication of supporting premises. Messages with explicitly articulated conclusions reduce audience misunderstanding and ensure the intended point is communicated (O'Keefe, 1997). For lower-involvement audiences, premise density additionally functions as a peripheral "length-is-strength" heuristic under ELM (Petty and Cacioppo, 1984).

SPECIFICITY: linguistic abstraction and psychological distance of the argument's subject. Construal Level Theory argues that concrete, specific language is more persuasive for immediate targets as it facilitates perceptual vividness and increases perceived truth (Fujita et al., 2008; Hansen and Wänke, 2010; Packard and Berger, 2021). Interactional immediacy via second-person pronoun use further triggers self-referencing, deepening receiver

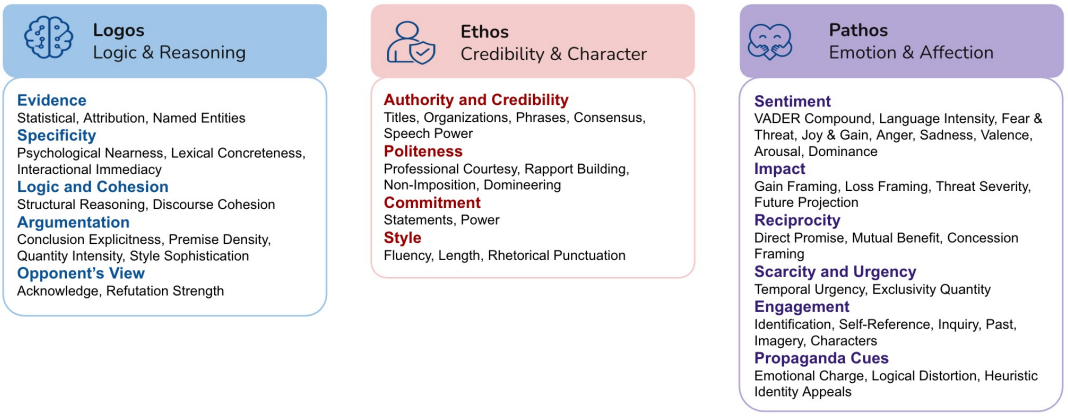

Figure 2: The Persuasion Index (PI) taxonomy. Fifteen dimensions are organized under the Aristotelian triad of **Logos**, **Ethos**, and **Pathos**, with sub-features listed under each dimension.

involvement (Cruz et al., 2017).

OPPONENT'S VIEW: acknowledgment and refutation of opposing positions. Refutational two-sided messages, which acknowledge then attack the opposing position, are more persuasive than one-sided messages because they preemptively resolve receiver objections (O'Keefe, 1999). Non-refutational two-sided messages are consistently less effective than one-sided appeals (Allen, 1991).

3.2 Ethos: The Appeal to Character

Ethos captures credibility signals embedded in linguistic choices, functioning as heuristic shortcuts that shape receiver trust independently of argument content (Hovland and Weiss, 1951; Chaiken, 1980).

AUTHORITY AND CREDIBILITY: expertise and institutional trust. High-credibility markers generally enhance persuasion, though their influence is dynamic and contingent on receiver involvement (Pornpitakpan, 2004). The PI also considers trust-building markers to counteract the psychological reactance that excessive authority distance can trigger (Hovland and Weiss, 1951; Song et al., 2018).

POLITENESS: how an argument mitigates imposition on the receiver, reducing psychological reactance. Successful communicators mitigate the "Face Threatening Acts" inherent in persuasive requests (Brown and Levinson, 1987). Meta-analyses of the "But You Are Free" technique demonstrate that acknowledging a listener's autonomy significantly increases compliance by reducing the perceived threat to freedom (Carpenter, 2013).

COMMITMENT: communicator's resolve and investment. The "Foot-in-the-Door" mechanism shows that expressed prior effort shifts a receiver's perception of the speaker's consistency (Burger, 1999). Persuasive effects are stronger when there is verifiable proof of action (Pallak et al., 1980).

STYLE: surface-level markers of communicative competence that influence the text's perceptual fluency. Individuals often mistake the ease of processing a text for factual accuracy, using fluency as a heuristic for truth (Reber and Schwarz, 1999). Terminal punctuation and message length function as social cues for sincerity and emotional intensity in digital communication (Kemp et al., 2025).

3.3 Pathos: The Appeal to Emotion

Pathos captures affective and social compliance triggers, operating primarily through peripheral routes where emotional resonance influences judgment without requiring deep logical scrutiny (Forgas, 1995; Petty and Cacioppo, 1986).

SENTIMENT (EMOTION): the affective intensity of language. Discrete emotions produce distinct action tendencies that influence processing: fear promotes protection-seeking while anger promotes heuristic evaluation (Petty et al., 1991; Schwarz and Clore, 1983). Meta-analyses on fear appeals suggest they are highly effective when threat is balanced with efficacy (Witte and Allen, 2000).

IMPACT: future consequences with projected gains or losses. The linguistic framing of costs and benefits significantly alters receivers' psychological cost-benefit analysis (Kahneman and Tversky, 1979; O'Keefe and Jensen, 2007). The persuasive force of these frames is further conditioned by projected outcome severity, as messages depicting high-severity consequences drive receivers toward danger-control processing (Witte and Allen, 2000).

ENGAGEMENT: textual antecedents hypothesized to stimulate psychological immersion (i.e., trans-

portation) or prompt audience self-referencing. While narrative transportation and identification represent subjective audience states rather than inherent message properties, they are systematically activated by distinct story-driven markers; audience transportation promotes persuasion by reducing the receiver's ability to counter-argue, bypassing logical scrutiny (Green and Brock, 2000; Slater and Rouner, 2002). Self-referencing effects further confirm that direct audience address integrates message content with the receiver's self-concept, making persuasive outcomes more persistent (Burnkrant and Unnava, 1995).

RECIPROCITY: concession or indebtedness prompting an obligation to comply. The "Door-in-the-Face" technique is driven by reciprocal concession mechanisms, where perceived social sacrifice triggers obligation in the receiver (O'Keefe and Hale, 1998). The "That's-Not-All" technique suggests that pre-emptive deal-sweetening creates immediate indebtedness (Burger, 1986).

SCARCITY & URGENCY: limited time or availability. Scarcity heightens perceived value and acts as a competitive signal that escalates decision urgency (Ladeira et al., 2023). Per Psychological Reactance Theory: when freedom of choice is threatened by limited supply or deadlines, individuals are driven to pursue the restricted object more intensely (Brehm, 1966; Worchel et al., 1975).

PROPAGANDA: identity-based heuristics and logical distortions that secure compliance regardless of logical depth. Even modest propaganda exposure produces significant shifts in pro-regime attitudes, though effects vary considerably across contexts (Shen and Yin, 2025). Propaganda further functions through deterrence, using markers of strength and danger to suppress counter-argumentation through fear (Skuhala Karasman, 2020; Jowett and O'Donnell, 2012).

4 Methods

We present a transparent implementation of PI using lexicon and rule-based sub-features. Crucially, the framework is implementation-agnostic: lexicons can be swapped for other models while preserving the theoretically-grounded PI structure. Section 4.3 describes our empirical setup.

4.1 Feature Construction and Validation

We represent each text as a fixed-dimensional vector corresponding to the PI sub-features. Table A2 contains specific operationalizations for all sub-features. We first leverage existing computational linguistic resources when available, including concreteness norms (Brysbaert et al., 2014) (SPECIFICITY), LIWC and NRC-VAD lexicons for SENTIMENT and ENGAGEMENT (Boyd et al., 2022; Mohammad, 2025), and VADER for polarity scoring (Hutto and Gilbert, 2014). We also use regular expressions to detect structural cues, e.g., numbers and URLs under EVIDENCE.

For remaining sub-features, we first manually constructed seed lexicons based on foundational persuasion theories (Table A1). For example, the `Causal` sub-feature of the LOGIC AND COHESION dimension has a seed lexicon containing causal connectives ("because", "therefore") drawn from argumentation theory (Toulmin, 2003). After manually establishing construct validity of these seed lexicons, we expand each one with a series of structured prompts to `GPT-5.4-mini` that specify the sub-feature label, theoretical anchor, inclusion/exclusion criteria, and one of seven register- or morphology-conditioned slices (e.g., formal, colloquial, domain-specific, inflectional variants). Generated candidates are retained or filtered out via a hybrid procedure of direct human annotation and human-validated LLM-as-judge. We test expanded lexicon robustness with a split-half consistency analysis and observe high correlations across the both lexicon-driven sub-features and mean-features. Full details are available in Appendix I.

4.2 PI Score Generation

Each sub-feature is scored in [0,1] as either a saturating density $1 - e^{-k \cdot r}$ (where r is the per-100-token match rate and k=0.5 controls saturation speed) or a binary indicator for sparse signals. The value k=0.5 is chosen so that r=2 (two matches per 100 tokens) maps to a moderate score of ≈ 0.63 and saturation approaches 1 for $r \geq 8$, calibrating the curve to the typical match density observed in our corpora. Overlapping lexicon matches are resolved by retaining only the longest non-contained span. Dimension scores are the mean of their sub-features, yielding a 15-dimensional category-mean vector or a full 55-dimensional sub-feature vector.

4.3 Evaluation Datasets and Models

We evaluate PI on four publicly released persuasion datasets (Appendix B): **UKPConvArg1 (UKP)** (Habernal and Gurevych, 2016) (n=11,650, pairwise convincingness), **CMV** (n=8,526, Δ-based

	PI-sub	PI-mean	RoBERTa	GPT-4o
UKP	0.768	0.684	0.565	**0.842**
CMV	**0.588**	0.571	0.459	0.492
IBM	**0.590**	0.579	0.225	0.586
Anthropic	0.432	0.429	0.400	**0.490**

Table 1: Persuasion prediction F1 scores. `PI-sub` uses 55 sub-features; `PI-mean` uses 15 mean-features. UKP/IBM use pairwise difference vectors; CMV/Anthropic use single-argument vectors. `RoBERTa` is fine-tuned; `GPT-4o` is zero-shot.

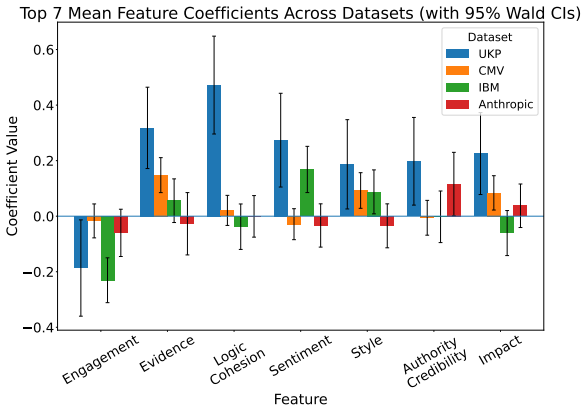

Figure 3: Top 7 `PI-mean` dimensions by mean absolute coefficient across datasets (95% Wald CI).

binary outcomes), **IBM Argument Quality** (Gretz et al., 2020) (n=14,003, expert quality ratings with pairwise comparisons), and **Anthropic Persuasion** (Durmus et al., 2024) (n=3,882, argument-level annotations). All datasets are normalized into a unified binary persuasion label. For pairwise datasets (UKP and IBM), we construct difference vectors $\Delta\mathbf{x} = \mathbf{x}_A - \mathbf{x}_B$ from each argument pair's PI feature vectors, with label 1 if argument A is judged more persuasive. For CMV, each reply is labeled 1 if the original poster awarded a Δ, and 0 otherwise. For Anthropic, we label 1 if post-exposure attitude improves by more than 1 point after removing participants with extreme initial attitudes due to ceiling and floor effects (Appendix C).

We compare ℓ_2-regularized logistic regression on PI feature vectors (15-dimensional category means or full 55-dimensional sub-features) against fine-tuned `RoBERTa` (Liu et al., 2019) and zero-shot `GPT-4o` as opaque high-capacity baselines. All models use an 80/20 stratified train–test split; for UKP and IBM, the split is applied at the argument level so that no base argument appears in both train and test across its pairings (Appendix D).

5 Results

Predictive Performance Our straightforward implementation of PI carries predictive signal for persuasion, with performance comparable to larger black-box models. `PI-sub` consistently outperforms `PI-mean` across all datasets (Table 1), confirming that preserving sub-feature granularity aids persuasion modeling.

`PI-sub` performs competitively with our black-box LLM baselines. On UKP, `PI-sub` reaches parity with the original benchmark (0.76–0.78) (Habernal and Gurevych, 2016), outperforming `RoBERTa` and trailing zero-shot `GPT-4o`. On CMV and IBM, `PI-sub` achieves the highest F1 among all models considered. `RoBERTa` underperforms across all four datasets and collapses on IBM (F1 = 0.225). Overall, `PI-sub` and `GPT-4o` each lead on two datasets, but `PI-sub` attains this with a transparent and computationally lightweight feature pipeline.

We statistically compare models via McNemar's test, bootstrap difference tests, and equivalence testing (TOST). `PI-sub` significantly outperforms `RoBERTa` on UKP, CMV, and IBM, and significantly outperforms `GPT-4o` on CMV and Anthropic, while `GPT-4o` significantly outperforms `PI-sub` on UKP. See full results and statistical tests in Appendix E.

These results show that PI reliably ranks argument persuasiveness across four datasets that vary substantially in domain, style, and outcome measures. However, we caution that these are simple LLM baselines: `RoBERTa` is fine-tuned with a fixed seed and no hyperparameter search, and `GPT-4o` is evaluated zero-shot with a minimal prompt. Rather than claim state-of-the-art performance, we treat these comparisons as evidence of predictive validity. PI provides a transparent feature space that is not only interpretable for descriptive analysis, but also predictive enough to support meaningful comparisons of rhetorical patterns across domains.

Coefficients by Dataset Figure 3 shows 7 PI dimensions with the largest mean coefficient magnitudes across datasets (coefficient robustness is tested via VIF and ablation analyses; Appendix H). UKP exhibits the largest and most consistently positive coefficients overall, led by LOGIC/COHESION, EVIDENCE, and SENTIMENT, patterns consistent with the structured, evidence-oriented nature of its crowdsourced pairwise convincingness task. LOGIC/COHESION is highly domain-specific: dominant on UKP but statistically indistinguishable from zero on CMV, IBM, and Anthropic. EVIDENCE

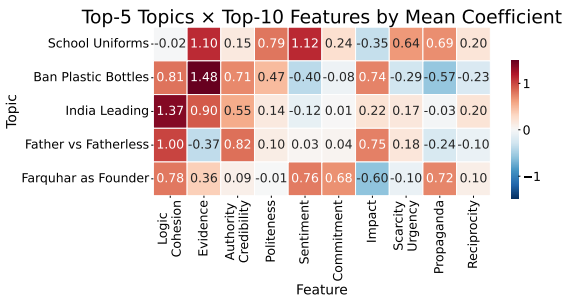

Figure 4: Top 10 `PI-mean` dimensions by mean absolute coefficient for the 5 highest-accuracy UKP topics. Cells show standardized coefficients.

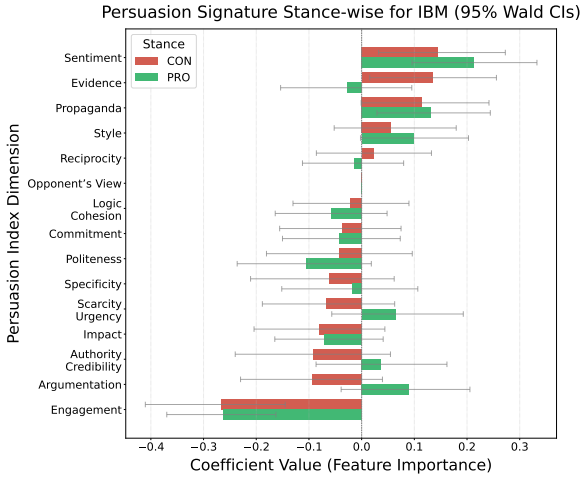

Figure 5: Coefficients for all 15 PI dimensions by stance in the IBM dataset (95% Wald CI). PRO (CON) arguments advocate (oppose) the stated premise.

generalizes more broadly, retaining a clearly positive effect on CMV but attenuating on IBM and Anthropic. Each remaining dataset contributes one distinctive signal of its own, SENTIMENT on IBM and AUTHORITY/CREDIBILITY on Anthropic, while their other coefficients are small with confidence intervals that frequently span zero. ENGAGEMENT has negative coefficients across all datasets, suggesting that personal pronoun density and narrative immersion do not reliably distinguish more persuasive arguments in these corpora.

Coefficients by Topic Figure 4 shows that persuasive strategies shift substantially across debate topics. EVIDENCE is the most consistent predictor, peaking in *Ban Plastic Bottles* (1.48), where LOGIC/COHESION (0.81) and AUTHORITY/CREDIBILITY (0.71) also rank highly, consistent with policy topics rewarding factual, structured reasoning. *India Leading* shows a similar Logos-dominant profile, with LOGIC/COHESION reaching its highest value (1.37) alongside strong EVIDENCE (0.90). Interpersonal topics shift toward affective and relational dimensions: *Father vs. Fatherless* is driven by LOGIC/COHESION (1.00) and IMPACT (0.75), while *Farquhar as Founder* relies on SENTIMENT (0.76) and COMMITMENT (0.68) with a notably negative IMPACT (−0.60). *School Uniforms* presents a distinct mix of EVIDENCE (1.10), SENTIMENT (1.12), and PROPAGANDA (0.69), reflecting a topic where normative pressure and emotional appeals dominate over logical structure. These patterns confirm that while EVIDENCE and LOGIC/COHESION provide stable cross-topic signals, the relative importance of credibility-based, affective, and persuasion-pressure dimensions varies substantially with topic framing. Full results across all 16 UKP topics and all features, as well as IBM topic-level results, are provided in Appendix F.1.

Coefficients by Stance Each topic in the IBM dataset has arguments on both sides: PRO advocates for a stated premise (e.g., *We should adopt vegetarianism*) while CON opposes it (e.g., *We should abandon vegetarianism*) (Appendix B). Models are fitted separately for each stance. Persuasion is stance-contingent in this setting (Figure 5). Both stances share SENTIMENT as the strongest positive predictor and ENGAGEMENT as the strongest negative predictor. PRO arguments also weight ARGUMENTATION, AUTHORITY/CREDIBILITY, and SCARCITY/URGENCY positively, suggesting that supporting a claim benefits from explicit reasoning, credibility appeals, and decision pressure. CON arguments weight EVIDENCE and RECIPROCITY more heavily, consistent with a strategy of factual challenge and concession framing when opposing a premise.

6 Discussion

Persuasion research faces a persistent tension between predictive performance and interpretability. Neural and LLM-based models achieve strong results but treat persuasion as a scalar outcome, offering no insight into which rhetorical mechanisms drive success. PI addresses this gap by grounding persuasion measurement in an explicit, theory-derived feature space, and implemented through transparent lexicons and rule-based detectors, where every score traces back to specific lexical or structural patterns in the text.

Empirical validation on four datasets indicates that PI features can capture meaningful predictive

signal for persuasion-related outcomes. PI-sub performs competitively, and sometimes outperforms, opaque RoBERTa and GPT-4o models. Unlike post-hoc methods that approximate feature importance after an opaque model has already predicted, PI is interpretable by construction: the features used by the predictive model are the same features visible to the researcher, and learned coefficients provide insights into associations between rhetorical dimensions and persuasion outcomes.

The most substantively important finding is not that PI predicts persuasion, but that comparative analyses across datasets, topics, and stances reveal regularities and variation in what kind of language tends to be persuasive. EVIDENCE, SENTIMENT, and LOGIC/COHESION are broadly positive predictors, but their relative weights vary across settings. UKP places more weight on logical structure and factual grounding; CMV shows more variable patterns, consistent with informal debate; Anthropic assigns higher weights to authority and credibility. EVIDENCE and ARGUMENTATION are linked to persuasion for policy topics, while SENTIMENT and IMPACT are stronger predictors for interpersonal topics. At the stance level, PRO arguments (supporting a given claim) have higher weights for ARGUMENTATION and AUTHORITY/CREDIBILITY, while CON arguments (opposing a given claim) rely on EVIDENCE and RECIPROCITY.

The consistently negative coefficient for ENGAGEMENT is notable. In these corpora, high personal pronoun density, direct address, and rhetorical questions do not reliably distinguish more persuasive arguments. However, this does not necessarily imply that narrative or engagement-oriented strategies are ineffective; rather, the surface linguistic markers we can extract from text may be weak proxies for audience-side psychological processes, such as identification or involvement. The effectiveness of engagement-oriented language likely depends on factors beyond the text itself, including audience prior attitudes, medium, and context.

These patterns demonstrate that persuasion is better treated as a context-sensitive rhetorical configuration than a single generalizable signal. From an ELM perspective, structured debate audiences may engage the central route more systematically, rewarding logical coherence, while informal settings elevate peripheral cues. PI makes these shifts quantifiable in an interpretable feature space and helps move NLP and computational social science persuasion research toward the kind of theoretical explanation that has long been central to psychology and communication scholarship.

PI opens several practical directions. It facilitates content analysis across domains. For social scientists who study persuasive message effects, it provides a principled framework for constructing stimuli and conducting automated manipulation checks. By verifying that only the targeted rhetorical dimensions vary across experimental conditions while others are held strictly constant, it helps researchers eliminate confounding variables—a level of precision difficult to achieve with manual text validation or with opaque neural systems. For researchers studying human–LLM or multi-agent persuasive dynamics, PI offers a shared feature space applicable to both human- and model-generated text, enabling direct rhetorical comparison across sources. As LLMs become increasingly capable persuaders (Hackenburg et al., 2025b; Breum et al., 2024), characterizing which rhetorical mechanisms they deploy and how these compare to human norms becomes critical for accountability.

PI is also designed to grow. Because the taxonomy and the implementation are decoupled, any individual dimension or sub-feature detector can be updated. For example, lexicons can be replaced with fine-tuned classifiers or discourse-aware models without disturbing the 15-dimension structure or its theoretical anchors. This separation distinguishes two questions that prior work tends to conflate: what rhetorical dimensions matter for persuasion (the taxonomy) and how best to detect each one in text (the implementation). PI offers a stable answer to the first while inviting community contributions to the second. To support this, we release PI as an open-source Python package and a public web interface for real-time argument analysis (Appendix J). We envision PI as a community-driven resource where researchers test the framework on new domains, swap in stronger detectors for individual sub-features, and extend the feature pipeline, progressively expanding its theoretical coverage and cross-domain generalizability.

The interpretability–performance tradeoff in persuasion modeling is smaller than often assumed. What PI sacrifices in raw predictive power, it recovers in theoretical legibility, domain portability, and practical utility, offering a foundation for persuasion analysis that is competitive empirically, auditable by design, and directly connected to theoretical constructs that make findings meaningful.

7 Limitations

A fundamental limitation of PI, and of text-based persuasion modeling more broadly, is that persuasiveness is not fully determined by message content alone. Relational context, source reputation, audience prior attitudes, and social dynamics all shape persuasive outcomes but are not recoverable from text. PI operationalizes persuasion as a property of the message as a linguistic artifact, capturing the rhetorical choices a communicator encodes in text but not how those choices interact with receiver characteristics or interactional context.

PI's lexicon- and rule-based implementations prioritize transparency but are limited to surface-level operationalization. Strategies that depend on pragmatic interpretation, such as implicit framing, irony, and sarcasm, as well as discourse-level phenomena such as long-range narrative structure and cross-sentence argumentation, remain underrepresented. PI's modular design partially mitigates this: individual sub-feature implementations can be replaced or augmented with richer methods, including neural or LLM-based detectors, without altering the taxonomic structure or its theoretical grounding.

Several dimensions show confidence intervals overlapping zero in some datasets, reflecting both genuine cross-context variation in persuasive norms and the noisiness of sparse lexicon coverage. More data would yield more precise coefficient estimates, and continued lexicon refinement remains necessary. Per-feature internal consistency results are reported in Table A18 as a calibrated disclosure of which dimensions provide robust signals and which require further development.

Finally, PI is evaluated exclusively on English-language argumentative text from online debate, crowdfunding, and controlled experiment settings. Generalization to other languages, cultural rhetorical norms, spoken discourse, or multimodal contexts remains an open question. Lexicons constructed from English argumentative conventions may systematically fail to capture persuasive strategies prevalent in other linguistic and cultural traditions. Extending PI to multilingual and multimodal settings is an important direction for future work.

8 Ethics Statement

The Persuasion Index is designed as a transparent analytical tool for studying persuasive language. However, a framework that identifies effective persuasion strategies could in principle be misused to engineer more persuasive misinformation, manipulative political messaging, or targeted influence campaigns. We emphasize that PI is intended for analysis and auditing of persuasive content rather than its generation. Its transparency is itself a safeguard: by making persuasion mechanisms explicit and traceable, PI enables detection and critique of manipulative rhetoric as much as it describes it. Lexicon-based features calibrated on English argumentative text may not generalize equally across languages, cultural communities, or demographic groups, as rhetorical norms vary substantially across contexts. We encourage users to audit PI scores for systematic disparities before deploying the framework in applied settings. All datasets used are publicly released and contain no personally identifiable information. Lexicon annotation involved only word-level judgments with no collection of personal data.

References

Mike Allen. 1991. Meta-analysis comparing the persuasiveness of one-sided and two-sided messages. *Western Journal of Speech Communication*, 55(4):390–404.

Mike Allen and Raymond W. Preiss. 1997. Comparing the persuasiveness of narrative and statistical evidence using meta-analysis. *Communication Research Reports*, 14(2):125–131.

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to Ask for a Favor: A Case Study on the Success of Altruistic Requests. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):12–21.

Aristotle. 2007. *On rhetoric: A theory of civic discourse*, 2nd ed. edition. Oxford University Press, New York.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Hui Bai, Jan G. Voelkel, Shane Muldowney, Johannes C. Eichstaedt, and Robb Willer. 2025. LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037.

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. 2024. Decoding persuasion: a survey on ML

and NLP methods for the study of online persuasion. *Frontiers in Communication*, 9.

Elisabeth Bigsby, Cabral A. Bigman, and Andrea Martinez Gonzalez. 2019. Exemplification theory: A review and meta-analysis of exemplar messages. *Annals of the International Communication Association*, 43(4):273–296.

Kevin L. Blankenship and Traci Y. Craig. 2006. Rhetorical Question Use and Resistance to Persuasion: An Attitude Strength Analysis. *Journal of Language and Social Psychology*, 25(2):111–128.

Jack Blumenau and Benjamin E. Lauderdale. 2024. The variable persuasiveness of political rhetoric. *American Journal of Political Science*, 68(1):255–270.

Ryan L Boyd, Ashwini Ashokkumar, Sara Seraj, and James W. Pennebaker. 2022. The Development and Psychometric Properties of LIWC-22.

Nimet Beyza Bozdag, Shuhaib Mehri, Xiaocheng Yang, Hyeonjeong Ha, Zirui Cheng, Esin Durmus, Jiaxuan You, Heng Ji, Gokhan Tur, and Dilek Hakkani-Tür. 2026. Must Read: A Comprehensive Survey of Computational Persuasion. *ACM Computing Surveys*, 58(12):1–39.

Jack W. Brehm. 1966. *A theory of psychological reactance*. Academic Press, Oxford, England.

Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The persuasive power of large language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:152–163.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge University Press.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3):904–911.

Jan-Willem H. Bullée, Lorena Montoya, Wolter Pieters, Marianne Junger, and Pieter H. Hartel. 2015. The persuasion and security awareness experiment: reducing the success of social engineering attacks. *Journal of Experimental Criminology*, 11(1):97–115.

Jerry M. Burger. 1986. Increasing compliance by improving the deal: The that's-not-all technique. *Journal of Personality and Social Psychology*, 51(2):277–283.

Jerry M. Burger. 1999. The foot-in-the-door compliance procedure: A multiple-process analysis and review. *Personality and Social Psychology Review*, 3(4):303–325.

Robert E. Burnkrant and H. Rao Unnava. 1995. Effects of self-referencing on persuasion. *Journal of Consumer Research*, 22(1):17–26.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

Christopher J. Carpenter. 2013. A meta-analysis of the effectiveness of the "but you are free" compliance-gaining technique. *Communication Studies*, 64(1):6–17.

Shelly Chaiken. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5):752–766.

Jiaao Chen and Diyi Yang. 2021. Weakly-Supervised Hierarchical Models for Predicting Persuasive Strategies in Good-faith Textual Requests. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12648–12656.

Junhan Chen, Shilin Xia, and Tong Lin. 2024. A framework of moderators in social norm-based message persuasiveness based on a systematic review. *Human Communication Research*, 50(2):285–298.

Zirui Cheng and Jiaxuan You. 2026. Towards Strategic Persuasion with Language Models. In *Proceedings of the International Conference on Learning Representations*. ArXiv:2509.22989 [cs].

David Ciuk and Joshua Rottman. 2020. Moral Conviction, Emotion, and the Influence of Episodic versus Thematic Frames. *Political Communication*, 38:1–20.

Thomas H. Costello, Gordon Pennycook, and David G. Rand. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714):eadq1814.

Ryan E. Cruz, James M. Leonhardt, and Todd Pezzuti. 2017. Second person pronouns enhance consumer involvement and brand attitude. *Journal of Interactive Marketing*, 39(1):104–116.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-Grained Analysis of Propaganda in News Articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages

5636–5646, Hong Kong, China. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Saloni Dash, Yiwei Xu, Madeline Jalbert, and Emma S Spiro. 2025. The persuasive potential of AI-paraphrased information at scale. *PNAS Nexus*, 4(7):pgaf207.

Alice Dechêne, Christoph Stahl, Jochim Hansen, and Michaela Wänke. 2010. The truth about the truth: a meta-analytic review of the truth effect. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, 14(2):238–257.

James P. Dillard. 2014. Language, style, and persuasion. In Thomas M. Holtgraves, editor, *The Oxford Handbook of Language and Social Psychology*, pages 177–187. Oxford University Press, New York, NY.

James Price Dillard and Lijiang Shen, editors. 2012. *The SAGE Handbook of Persuasion: Developments in Theory and Practice*, 2 edition. SAGE Publications, Inc., Thousand Oaks, CA.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 Task 6: Detection of Persuasion Techniques in Texts and Images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98. Association for Computational Linguistics.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. Measuring the Persuasiveness of Language Models.

Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pages 1–12, Uppsala, Sweden. Association for Computational Linguistics.

Joseph P. Forgas. 1995. Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin*, 117(1):39–66.

Kentaro Fujita, Tal Eyal, Shelly Chaiken, Yaacov Trope, and Nira Liberman. 2008. Influencing attitudes toward near and distant objects. *Journal of Experimental Social Psychology*, 44(3):562–572.

Kristel M. Gallagher and John A. Updegraff. 2012. Health message framing effects on attitudes, intentions, and behavior: a meta-analytic review. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine*, 43(1):101–116.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained Argumentation Features for Scoring Persuasive Essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany. Association for Computational Linguistics.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are You Convinced? Choosing the More Convincing Evidence with a Siamese Network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Naomi K. Grant, Leandre R. Fabrigar, Adelle Forzley, and Maia Kredentser. 2014. The multiple roles of scarcity in compliance: Elaboration as a moderator of scarcity mechanisms. *Social Influence*, 9(2):149–161.

Melanie C. Green and Timothy C. Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology*, 79(5):701–721.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7805–7813.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Kobi Hackenburg, Lujain Ibrahim, Ben M. Tappin, and Manos Tsakiris. 2025a. Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues. *AI & SOCIETY*.

Kobi Hackenburg and Helen Margetts. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2403116121.

Kobi Hackenburg, Ben M. Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G. Rand, and Christopher Summerfield. 2025b. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884.

Jochim Hansen and Michaela Wänke. 2010. Truth from language and truth from fit: The impact of linguistic concreteness and level of construal on subjective truth. *Personality & Social Psychology Bulletin*, 36(11):1576–88.

Christopher Hidey and Kathleen McKeown. 2018. Persuasive influence detection: The role of argument sequencing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the Semantic Types of Claims and Premises in an Online Persuasive Forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Colin Higgins and Robyn Walker. 2012. Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. *Accounting Forum*, 36(3):194–208.

Gia Bao Hoang, Keith J. Ransom, Rachel Stephens, Carolyn Semmler, Nicolas Fay, and Lewis Mitchell. 2025. A Hybrid Theory and Data-driven Approach to Persuasion Detection with Large Language Models. In *Proceedings of the International AAAI Conference on Web and Social Media Workshops*. AAAI Press.

Jacob Hornik, Chezy Ofir, and Matti Rachamim. 2016. Quantitative evaluation of persuasive appeals using comparative meta-analysis. *The Communication Review*, 19(3):192–222.

Carl I. Hovland and Wallace Mandell. 1952. An experimental comparison of conclusion-drawing by the communicator and by the audience. *The Journal of Abnormal and Social Psychology*, 47(3):581–588.

Carl I. Hovland and Walter Weiss. 1951. The influence of source credibility on communication effectiveness. *The Public Opinion Quarterly*, 15(4):635–650.

C. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Omkar Joshi, Priya Pitre, and Yashodhara Haribhakta. 2023. ArgAnalysis35K : A large-scale dataset for Argument Quality Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13916–13931, Toronto, Canada. Association for Computational Linguistics.

Garth S. Jowett and Victoria O'Donnell. 2012. *Propaganda & persuasion*, 5th ed edition. SAGE, Thousand Oaks, Calif.

Johanna K. Kaakinen, Jonna Salonen, Paula Venäläinen, and Jukka Hyönä. 2011. Influence of text cohesion on the persuasive power of expository text. *Scandinavian Journal of Psychology*, 52(3):201–208.

Daniel Kahneman and Amos Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291.

Nenagh Kemp, Rebecca Kovacic, and Elisabeth Beyersmann. 2025. Is the period really "pissed"? The effect of punctuation and message length on perceptions in digital communication. *Telematics and Informatics*, 97:102241.

Hye Kyung Kim and Minyi Chen. 2025. A meta-analysis of gain–loss framing effects in narrative persuasion. *Journal of Communication*, 75(3):220–230.

Sang-Yeon Kim, Mike Allen, Ali Gattoni, Denis Grimes, Anna M. Herrman, Han Huang, Jihyun Kim, Shan Lu, Melissa Maier, Amy May, Kim Omachinski, Kikuko Omori, Kelly Tenzek, Katie LaPlant Turkiewicz, and Yan Zhang. 2012. Testing an additive model for the effectiveness of evidence on the persuasiveness of a message. *Social Influence*, 7(2):65–77.

Wagner Ladeira, Weng Marc Lim, Fernando Santini, Tareq Rasul, Marcelo Perin, and Levent Altinay. 2023. A meta-analysis on the effects of product scarcity. *Psychology & Marketing*, 40:1267–1279.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. ArXiv:1907.11692 [cs].

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics.

Kelvin Luu, Chenhao Tan, and Noah A. Smith. 2019. Measuring Online Debaters' Persuasive Skill from Text over Time. *Transactions of the Association for Computational Linguistics*, 7:537–550. Place: Cambridge, MA.

S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. 2024. The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14(1):4692.

James C. McCroskey and Jason J. Teven. 1999. Goodwill: A reexamination of the construct and its measurement. *Communication Monographs*, 66(1):90–103.

James C. McCroskey and Thomas J. Young. 1981. Ethos and credibility: The construct and its measurement after three decades. *Central States Speech Journal*, 32(1):24–34.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Tanushree Mitra and Eric Gilbert. 2014. The language that gets people to give: phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, CSCW '14, pages 49–61, New York, NY, USA. Association for Computing Machinery.

Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. 2025. PCoT: Persuasion-Augmented Chain of Thought for Detecting Fake News and Social Media Disinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24959–24983, Vienna, Austria. Association for Computational Linguistics.

Saif Mohammad. 2018. Word Affect Intensities. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif M. Mohammad. 2025. NRC VAD Lexicon v2: Norms for Valence, Arousal, and Dominance for over 55k English Terms. *arXiv preprint*. ArXiv:2503.23547 [cs].

Corrado Monti, Luca Maria Aiello, Gianmarco De Francisci Morales, and Francesco Bonchi. 2022. The language of opinion change on social media under the lens of communicative action. *Scientific Reports*, 12(1):17920.

Emiko J. Muraki, Summer Abdalla, Marc Brysbaert, and Penny M. Pexman. 2023. Concreteness ratings for 62,000 English multiword expressions. *Behavior Research Methods*, 55(5):2522–2531.

Robin L. Nabi. 2002. Discrete Emotions and Persuasion. In *The Persuasion Handbook: Developments in Theory and Practice*, pages 289–308. SAGE Publications, Inc.

Daniel J. O'Keefe. 1997. Standpoint explicitness and persuasive effect: A meta-analytic review of the effects of varying conclusion articulation in persuasive messages. *Argumentation and Advocacy*, 34(1):1–12.

Daniel J. O'Keefe. 2015. *Persuasion: Theory and Research*, 3 edition. SAGE Publications, Thousand Oaks, CA.

Daniel J. O'Keefe and Jakob D. Jensen. 2007. The relative persuasiveness of gain-framed and loss-framed messages for encouraging disease prevention behaviors: a meta-analytic review. *Journal of Health Communication*, 12(7):623–644.

Daniel J. O'Keefe. 1999. How to handle opposing arguments in persuasive messages: A meta-analytic review of the effects of one-sided and two-sided messages. *Annals of the International Communication Association*, 22(1):209–249.

Daniel J. O'Keefe and Scott L. Hale. 1998. The door-in-the-face influence strategy: A random-effects meta-analytic review. *Annals of the International Communication Association*, 21(1):1–33.

Grant Packard and Jonah Berger. 2021. How concrete language shapes customer satisfaction. *Journal of Consumer Research*, 47(5):787–806.

Michael S. Pallak, David A. Cook, and John J. Sullivan. 1980. Commitment and energy conservation. In Leonard Bickman, editor, *Applied Social Psychology Annual*, volume 1, pages 235–253. Sage Publications, US.

Amalie Brogaard Pauli, Isabelle Augenstein, and Ira Assent. 2025. Measuring and Benchmarking Large Language Models' Capabilities to Generate Persuasive Language. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10056–10075, Albuquerque, New Mexico. Association for Computational Linguistics.

Richard E. Petty and John T. Cacioppo. 1984. The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, 46(1):69–81.

Richard E. Petty and John T. Cacioppo. 1986. The elaboration likelihood model of persuasion. In Leonard Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 19, pages 123–205. Academic Press.

Richard E. Petty, Faith Gleicher, and Sara M. Baker. 1991. Multiple roles for affect in persuasion. In *Emotion and social judgments*, International series in experimental social psychology, pages 181–200. Pergamon Press, Elmsford, NY, US.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Chanthika Pornpitakpan. 2004. The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology*, 34(2):243–281.

Stephen A. Rains. 2013. The Nature of Psychological Reactance Revisited: a Meta-Analytic Review. *Human Communication Research*, 39(1):47–73.

Rolf Reber and Norbert Schwarz. 1999. Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, 8(3):338–342.

Roman Rietsche, Andrew Caines, Cornelius Schramm, Dominik Pfütze, and Paula Buttery. 2022. The Specificity and Helpfulness of Peer-to-Peer Feedback in Higher Education. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 107–117, Seattle, Washington. Association for Computational Linguistics.

Julia Romberg, Maximilian Maurer, Henning Wachsmuth, and Gabriella Lapesa. 2025. Towards a Perspectivist Turn in Argument Quality Assessment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7458–7485, Albuquerque, New Mexico. Association for Computational Linguistics.

Alexander J. Rothman, Roger D. Bartels, Jhon Wlaschin, and Peter Salovey. 2006. The Strategic Use of Gain- and Loss-Framed Messages to Promote Healthy Behavior: How Theory Can Inform Practice. *Journal of Communication*, 56(Suppl 1):S202–S220.

Ahmed Sajwani, Alaa El Setohy, Ali Mekky, Diana Turmakhan, Lara Hassan, Mohamed El Zeftawy, Omar El Herraoui, Osama Mohammed Afzal, Qisheng Liao, Tarek Mahmoud, Zain Muhammad Mujahid, Muhammad Umar Salman, Muhammad Arslan Manzoor, Massa Baali, Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2024. FRAPPE: FRAming, Persuasion, and Propaganda Explorer. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 207–213, St. Julians, Malta. Association for Computational Linguistics.

Francesco Salvi, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2025. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 9(8):1645–1653.

Norbert Schwarz and Gerald L. Clore. 1983. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45(3):513–523.

Resyl Sebalda, Celeste Faye Taray, and Donnie M. Tulud. 2025. Analyzing coherence and cohesion in Sara Duterte's vice presidential inaugural speech. *International Journal of Language and Literary Studies*, 7(2):260–282.

Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effect of LLM-Powered Search Systems on Diverse Information Seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, ACM Conferences, pages 1–17.

Jocelyn Shen, Joel Mire, Hae Won Park, Cynthia Breazeal, and Maarten Sap. 2024. HEART-felt Narratives: Tracing Empathy and Narrative Style in Personal Stories with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1026–1046, Miami, Florida, USA. Association for Computational Linguistics.

Shuyuan Shen and Tianhong Yin. 2025. Does authoritarian propaganda persuade and deter? A meta-analysis of 70 Chinese propaganda experiments.

Ivana Skuhala Karasman. 2020. Propaganda as a mechanism of manipulation and encouragement to action. *RIThink*, 9.

Michael D. Slater and Donna Rouner. 2002. Entertainment—education and elaboration likelihood: understanding the processing of narrative persuasion. *Communication Theory*, 12(2):173–191.

Victoria Smith, Susan Siltanen, and Lawrence Hosman. 1998. The Effects of Powerful and Powerless Speech Styles and Speaker Expertise On Impression Formation and Attitude Change. *Comunication Research Reports*, 15(1):27–35.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting Arguing and Sentiment in Meetings. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 26–34, Antwerp, Belgium. Association for Computational Linguistics.

Hwanseok Song, Katherine McComas, and Krysten Schuler. 2018. Source effects on psychological reactance to regulatory policies: The role of trust and similarity. *Science Communication*, 40(4):525–553.

Christian Stab and Iryna Gurevych. 2014a. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014b. Identifying Argumentative Discourse Structures in Persuasive Essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Bagyasree Sudharsan and Maria Leonor Pacheco. 2025. Cross-Domain Persuasion Detection with Argumentative Features. In *Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 372–380, Suzhou, China. Association for Computational Linguistics.

Mark W. Susmann and Duane T. Wegener. 2023. The independent effects of source expertise and trustworthiness on retraction believability: The moderating role of vested interest. *Memory & Cognition*, 51(4):845–861.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 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 '16, pages 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic Argument Quality Assessment - New Datasets and Methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Stephen E. Toulmin. 2003. *The uses of argument*, 2 edition. Cambridge University Press, Cambridge.

Yaacov Trope and Nira Liberman. 2010. Construal-Level Theory of Psychological Distance. *Psychological review*, 117(2):440–463.

John Paul Vargheese, Matthew Collinson, and Judith Masthoff. 2020. Exploring Susceptibility Measures to Persuasion. In *Persuasive Technology. Designing for Future Change*, pages 16–29, Cham. Springer International Publishing.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational Argumentation Quality Assessment in Natural Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes. *Transactions of the Association for Computational Linguistics*, 5:219–232. Place: Cambridge, MA.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

K. Witte and M. Allen. 2000. A meta-analysis of fear appeals: implications for effective public health campaigns. *Health Education & Behavior*, 27(5):591–615.

Stephen Worchel, Jerry Lee, and Akanbi Adewole. 1975. Effects of supply and demand on ratings of object value. *Journal of Personality and Social Psychology*, 32(5):906–914.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let's Make Your Request More Persuasive: Modeling Persuasive Strategies via Semi-Supervised Neural Nets on Crowdfunding Platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics.

W. J. Youden. 1950. Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Simon Zebregs, Bas van den Putte, Peter Neijens, and Anneke de Graaf. 2015. The differential impact of statistical and narrative evidence on beliefs, attitude, and intention: A meta-analysis. *Health Communication*, 30(3):282–289.

Xiaoquan Zhao, Andrew Strasser, Joseph N. Cappella, Caryn Lerman, and Martin Fishbein. 2011. A Measure of Perceived Argument Strength: Reliability and Validity. *Communication methods and measures*, 5(1):48–75.

A Persuasion Strategies in the Persuasion Index

Table A1 provides the canonical reference for all 15 PI dimensions cited throughout the paper, organized under Aristotle's three modes of appeal, including Logos, Ethos, and Pathos. Each row lists the dimension's working definition with an illustrative example, the prior computational work informing its operationalization, and the theoretical literature grounding the underlying persuasive mechanism.

Table A2 maps each PI dimension to its constituent subfeatures and the specific linguistic cues used to detect them. The final column gives the regex, lexicon, or rule that produces each measurement; a subfeature score aggregates all of its assigned cues, normalized to per-token density unless the cue is marked `(binary)`, in which case the feature is computed as a 0/1 presence indicator per document. Together the two tables provide a complete trace from each Aristotelian appeal down to the operational rule that yields a feature value.

Strategy	Definition and Example	Computational Work	Theory Support
Evidence	Quantifies verifiable facts, statistics, and attributed information used to ground claims. *"Official statistics show unemployment fell by 20% this year."*	Gretz et al. 2020 Habernal and Gurevych 2016 Toledo et al. 2019 Min et al. 2023	Zebregs et al. 2015 Allen and Preiss 1997 Kim et al. 2012 Bigsby et al. 2019
Logic / Cohesion	Measures reasoning structure and discourse connectivity that reduce cognitive load in processing. *"Because the rate is fixed, your monthly payment will not increase."*	Hidey and McKeown 2018 Wang et al. 2017	Petty and Cacioppo 1986 Chaiken 1980 Kaakinen et al. 2011 Sebalda et al. 2025 Dechêne et al. 2010
Argumentation	Captures structural explicitness of claims, premises, and argumentative intensity within a text. *"Our city should expand transit because commute times doubled."*	Somasundaran et al. 2007 Hidey et al. 2017 Gretz et al. 2020 Joshi et al. 2023 Wachsmuth et al. 2017 Stab and Gurevych 2014b	Lukin et al. 2017 Blumenau and Lauderdale 2024 O'Keefe 1997 Hovland and Mandell 1952 Petty and Cacioppo 1984 Dillard 2014 Zhao et al. 2011
Specificity	Analyzes linguistic concreteness and psychological proximity to increase perceived truth. *"I contributed $25 last Friday so the shelter can vaccinate three more puppies."*	Althoff et al. 2014 Mitra and Gilbert 2014 Ghosh et al. 2016 Carlile et al. 2018 Brysbaert et al. 2014 Rietsche et al. 2022 Muraki et al. 2023	Trope and Liberman 2010 Hansen and Wänke 2010 Dillard 2014 Ciuk and Rottman 2020 Fujita et al. 2008 Cruz et al. 2017 Packard and Berger 2021
Opponent's View	Detects concession-refutation sequences that strategically acknowledge and rebut counterarguments. *"Some argue rent control reduces supply, but recent data disagree."*	Tan et al. 2016 Wang et al. 2017 Luu et al. 2019	O'Keefe 1999 Allen 1991
Authority / Credibility	Captures linguistic signals of expertise and institutional trust that function as credibility heuristics. *"Researchers at the national institute recommend this booster."*	Susmann and Wegener 2023 Vargheese et al. 2020 Dillard and Shen 2012 Farkas et al. 2010	Hovland and Weiss 1951 Dillard 2014 Song et al. 2018 Pornpitakpan 2004 Bullée et al. 2015 Smith et al. 1998 McCroskey and Young 1981 McCroskey and Teven 1999
Politeness	Measures face-saving language that reduces psychological reactance and mitigates persuasive imposition. *"I see your point, but may I suggest an alternative?"*	Danescu-Niculescu-Mizil et al. 2013 Tan et al. 2016 Luu et al. 2019 Farkas et al. 2010	Brown and Levinson 1987 Carpenter 2013 Rains 2013
Commitment	Signals the communicator's resolve and prior investment to establish a consistent, credible persona. *"I already pledged to bike to work, so I support this proposal."*	Vargheese et al. 2020 Yang et al. 2019 Wang et al. 2019	Burger 1999 Pallak et al. 1980 Chen et al. 2024
Surface / Style	Captures perceptual fluency and surface-level competence cues that serve as heuristics for truth. *"We will act, listen, and adapt."*	Ghosh et al. 2016 Wang et al. 2017 Gretz et al. 2020 Farkas et al. 2010	Kemp et al. 2025 Reber and Schwarz 1999
Sentiment	Measures affective intensity through polarity, discrete emotional categories, and dimensional affect scores. *"It is heartbreaking that so many families cannot afford care."*	Somasundaran et al. 2007 Monti et al. 2022 Breum et al. 2024 Mohammad 2018	Dillard 2014 Nabi 2002 Witte and Allen 2000 Hornik et al. 2016
Impact	Leverages Prospect Theory by framing projected gains, losses, and future consequences. *"Your donation funds three counseling sessions."*	Monti et al. 2022 Costello et al. 2024 Matz et al. 2024 Bai et al. 2025.	O'Keefe and Jensen 2007 Kim and Chen 2025 Witte and Allen 2000 Rothman et al. 2006 Dillard 2014 Carpenter 2013 Gallagher and Updegraff 2012

Continued on next page

Table A1 – *continued from previous page*

Strategy	Definition and Example	Computational Work	Theory Support
Engagement	Promotes narrative transportation via direct address, self-referencing, and vivid imagery. *"Have you ever wondered where your taxes go?"*	Tan et al. 2016 Luu et al. 2019	Dillard 2014 Green and Brock 2000 Burger 1999 Shen et al. 2024 Burnkrant and Unnava 1995 Blankenship and Craig 2006 Zebregs et al. 2015
Reciprocity	Triggers the norm of social exchange by signaling concession or indebtedness to prompt compliance. *"I have been sharing my notes all semester—could you sign this petition?"*	Vargheese et al. 2020 Althoff et al. 2014 Mitra and Gilbert 2014 Yang et al. 2019 O'Keefe and Hale 1998 Burger 1986	
Scarcity / Urgency	Heightens perceived value by invoking time limits or limited availability to intensify decision pressure. *"Only 24 hours remain before matching funds expire."*	Vargheese et al. 2020 Dillard and Shen 2012	Ladeira et al. 2023 Grant et al. 2014
Propaganda	Detects identity-based heuristics and logical distortions that secure compliance regardless of logical depth. *"Everyone who cares about freedom is joining."*	Dimitrov et al. 2021 Da San Martino et al. 2020 Piskorski et al. 2023 Modzelewski et al. 2025	Witte and Allen 2000 Shen and Yin 2025 Skuhala Karasman 2020

Table A1: Fifteen persuasion dimensions in the Persuasion Index (PI), organized under the Aristotelian triad. **Logos** (reasoning and evidence): Evidence, Logic/Cohesion, Argumentation, Specificity, Opponent's View. **Ethos** (credibility and character): Authority/Credibility, Politeness, Commitment, Style. **Pathos** (emotion and compliance): Sentiment, Impact, Engagement, Reciprocity, Scarcity/Urgency, Propaganda. Each row lists the dimension definition with an illustrative example, prior computational work that informs the operationalization, and the theoretical literature grounding the persuasive mechanism.

Appeal	Dimension	Subfeature	Linguistic Cue	Operationalization
Logos	Evidence	statistical	Numeric tokens	Regex RE_NUMBER: \\b\\d[\\d,.\\%]*\\b
Logos	Evidence	statistical	Percentage expressions	Regex RE_PERCENT: \\b\\d+%\\b; \\b\\d+\\spercent\\b
Logos	Evidence	statistical	Measurement/statistical units	Lexicon EVI_UNITS: kg, percent, million
Logos	Evidence	attribution	APA-style citations	Regex RE_APA (binary): [Author] (YYYY)
Logos	Evidence	attribution	Bracketed numeric citations	Regex RE_NUM_CIT (binary): [1]; [1, 2]
Logos	Evidence	attribution	URLs/web references	Regex RE_URL (binary): https?://\\S+; www.\\S+
Logos	Evidence	attribution	Authority/source phrases	Lexicon AUTHORITY_PHRASE (binary): according to, studies show, data from
Logos	Evidence	named_entities	Named entities as evidential anchors	spaCy NER: PERSON, ORG, GPE
Logos	Specificity	psychological_nearness	Explicit date/time cues	Regex RE_DATE_MD (binary): Jan 1, Sep. 12, 2020; Regex RE_TIME_SPECIFIC (binary): 10:30, morning, pm
Logos	Specificity	psychological_nearness	Psychologically near anchors	Lexicon SPEC_PSYCH_NEAR (binary): minute, hour, day
Logos	Specificity	psychological_nearness	Psychologically far anchors	Lexicon SPEC_PSYCH_FAR (binary): year, decade, century
Logos	Specificity	psychological_nearness	Specificity anchors	Lexicon SPEC_ANCHORS (binary): specifically, namely, instance
Logos	Specificity	psychological_nearness	Vague quantifiers/adverbs	Lexicon SPEC_VAGUE (binary): many, some, most
Logos	Specificity	lexical_concreteness	Brysbaert concreteness ratings for single words	Resource Brysbaert_concreteness_dataset.xlsx: Word, Conc.M
Logos	Specificity	lexical_concreteness	Multiword concreteness ratings	Resource MultiwordExpression_Concreteness_Ratings.csv: Expression, Mean_C
Logos	Specificity	interactional_immediacy	Second-person address	LIWC You: You
Logos	Specificity	interactional_immediacy	Person names near second-person address	spaCy PERSON proximity rule: PERSON, you, your
Logos	Logic/Cohesion	structural_reasoning	Causal connectives	Lexicon LOGIC_CAUSAL: because, since, as a result
Logos	Logic/Cohesion	structural_reasoning	Inferential cues	Lexicon LOGIC_INFERENCE: implies, suggests, indicates
Logos	Logic/Cohesion	structural_reasoning	Referential cohesion cues	Lexicon LOGIC_REFERENCE: this, that, these
Logos	Logic/Cohesion	structural_reasoning	If-then structures	Regex IF_THEN: \\bif\\b.*\\bthen\\b
Logos	Logic/Cohesion	discourse_cohesion	Contrastive discourse markers	Lexicon LOGIC_CONTRAST: however, although, though
Logos	Logic/Cohesion	discourse_cohesion	Additive/sequencing markers	Lexicon LOGIC_ADDITIVE: moreover, furthermore, in addition
Logos	Logic/Cohesion	discourse_cohesion	Repeated non-trivial content words	Regex repeated_content_words: \\b\\w{4,}\\b
Logos	Argumentation	conclusion_explicitness	Claim markers	Lexicon ARG_CLAIM: should, must, need to
Logos	Argumentation	conclusion_explicitness	Conclusion markers	Lexicon ARG_CONCLUSION: in conclusion, to sum up, therefore
Logos	Argumentation	premise_density	Premise/support markers	Lexicon ARG_PREMISE (binary): because, since, given that
Logos	Argumentation	quantity_intensity	Argumentative sentences containing claim/conclusion/premise markers	Sentence rule ARGUMENTATIVE_SENTENCES: ARG_CLAIM, ARG_CONCLUSION, ARG_PREMISE
Logos	Argumentation	style_sophistication	Argument style: definition cues	Lexicon ARG_STYLE.definition: means, refers to, defined as
Logos	Argumentation	style_sophistication	Argument style: analogy cues	Lexicon ARG_STYLE.analogy: like, as if, as though
Logos	Argumentation	style_sophistication	Argument style: example cues	Lexicon ARG_STYLE.example: for example, for instance, such as
Logos	Argumentation	style_sophistication	Argument style: cause_effect cues	Lexicon ARG_STYLE.cause_effect: because, therefore, thus
Logos	Argumentation	style_sophistication	Argument style: counter cues	Lexicon ARG_STYLE.counter: however, although, though
Logos	Argumentation	style_sophistication	Argument style: concession_refute cues	Lexicon ARG_STYLE.concession_refute: i agree that, you're right that, but
Logos	Argumentation	style_sophistication	Argument style: conditional cues	Lexicon ARG_STYLE.conditional: if, unless, provided that
Logos	Opponent's View	acknowledge	Opponent-view acknowledgment phrases	Lexicon OPP_VIEW (binary): i understand your point, i see your point, i agree that
Logos	Opponent's View	acknowledge	Soft acknowledgment frames	Regex RE_SOFT_ACK (binary): current approach, existing approach, limitations with the current approach
Logos	Opponent's View	refutation_strength	Acknowledgment followed by contrastive refutation	Sequential rule OPP_VIEW -> LOGIC_CONTRAST: i see your point -> however; i agree that -> but; you may think -> nevertheless
Ethos	Authority/Credibility	titles	Professional/academic titles	Regex RE_TITLES: dr., professor, phd
Ethos	Authority/Credibility	organizations	Institutional authority names	Regex RE_AUTH_ORG: world health organization, harvard university, cdc
Ethos	Authority/Credibility	organizations	Organization named entities	spaCy ORG NER: ORG
Ethos	Authority/Credibility	phrases	Authority/source phrases	Lexicon AUTHORITY_PHRASE (binary): according to, studies show, research indicates
Ethos	Authority/Credibility	consensus	Consensus/majority cues	Regex RE_CONSENSUS (binary): majority of, most people, nearly everyone
Ethos	Authority/Credibility	speech_power	Hedges as powerless speech cues	Lexicon HEDGES: may, might, can
Ethos	Authority/Credibility	speech_power	Hesitation/filler cues	Lexicon HESITATIONS: um, uh, err

Continued on next page

Table A2 – *continued from previous page*

Appeal	Dimension	Subfeature	Linguistic Cue	Operationalization
Ethos	Authority/Credibility	speech_power	Tag questions	Lexicon AUT_TAG_QUESTION: right?, don't you think?, isn't it?
Ethos	Politeness	professional_courtesy	Request politeness markers	Lexicon PLEASE (binary): please, plz, kindly
Ethos	Politeness	professional_courtesy	Gratitude markers	Lexicon THANKS (binary): thanks, thank you, appreciated
Ethos	Politeness	professional_courtesy	Formal courtesy terms	Lexicon COURTESY (binary): sir, ma'am, mr.
Ethos	Politeness	professional_courtesy	Apology/reassurance/greeting cues	Lexicon Apology (binary): sorry, woops, oops; Lexicon Reassurance (binary): is okay, is fine, you are good; Lexicon Greeting (binary): hi, hello, hey
Ethos	Politeness	rapport_building	Rapport-building phrases	Lexicon RAPPORT (binary): good point, well said, totally agree
Ethos	Politeness	rapport_building	Conversational rapport markers	Lexicon Greeting (binary): hi, hello, hey; Lexicon Filler (binary): um, uh, hmm; Lexicon By.The.Way (binary): by the way, btw; Lexicon Actually (binary): really, actually, honestly
Ethos	Politeness	non_imposition	Hedges/softeners	Lexicon HEDGES (binary): may, might, can
Ethos	Politeness	non_imposition	Indirect request forms	Lexicon INDIRECTS (binary): would you, could you, would it be possible
Ethos	Politeness	non_imposition	Subjunctive request forms	Lexicon Subjunctive (binary): could you, would you, could u
Ethos	Politeness	non_imposition	Indicative request forms	Lexicon Indicative (binary): can you, will you, can u
Ethos	Politeness	non_imposition	Perspective softeners	Lexicon For.Me (binary): for me, for us; Lexicon For.You (binary): for you, for u
Ethos	Politeness	domineering	Directive/command cues	Lexicon DOMINEERING (binary): must, have to, need to
Ethos	Commitment	statements	Personal/group commitment statements	Lexicon COMMITMENT (binary): i have already, i've already, i have been
Ethos	Commitment	power	Commitment followed by proof	Sequential rule COMMITMENT -> COMMITMENT_PROOF: i am committed -> provide documentation; we have -> show proof; together -> can verify
Ethos	Style	fluency	Out-of-vocabulary/low-frequency words	wordfreq Zipf rule: <2.5
Ethos	Style	length	Argument length	Formula LENGTH: log1p(token_count)/7
Ethos	Style	rhetorical_punctuation	Quotation marks/apostrophe-like punctuation	Regex quote_marks: ", “, ”
Ethos	Style	rhetorical_punctuation	Question marks	Regex question_marks: ?
Ethos	Style	rhetorical_punctuation	Exclamation marks	Regex exclamation_marks: !
Pathos	Sentiment	vader_compound	VADER compound polarity	Resource VADER: compound
Pathos	Sentiment	language_intensity	Extreme intensity words	Lexicon SENT_INTENSITY.extreme: catastrophic, devastating, horrific
Pathos	Sentiment	language_intensity	Strong intensity words	Lexicon SENT_INTENSITY.strong: absolutely, extremely, critical
Pathos	Sentiment	language_intensity	Moderate intensity words	Lexicon SENT_INTENSITY.moderate: significant, substantial, considerable
Pathos	Sentiment	language_intensity	Weak intensifiers	Lexicon SENT_INTENSITY.weak: really, very, quite
Pathos	Sentiment	language_intensity	Directive/urgency force cues	Weighted lexicons DOMINEERING/URGENCY_TIME/ARG_CLAIM/IMPACT_THREAT: must, urgent, should, catastrophic
Pathos	Sentiment	fear_threat	LIWC anxiety words	LIWC Anx: Anx
Pathos	Sentiment	fear_threat	Fear/threat lexicons	Lexicon PROP_FEAR: threat, danger, destroy; Lexicon IMPACT_THREAT: catastrophic, deadly, fatal
Pathos	Sentiment	joy_gain	Positive emotion words	LIWC Posemo: Posemo
Pathos	Sentiment	anger	Anger words	LIWC Anger: Anger
Pathos	Sentiment	sadness	Sadness words	LIWC Sad: Sad
Pathos	Sentiment	valence	NRC-VAD valence ratings	Resource NRC-VAD valence: valence
Pathos	Sentiment	arousal	NRC-VAD arousal ratings	Resource NRC-VAD arousal: arousal
Pathos	Sentiment	dominance	NRC-VAD dominance ratings	Resource NRC-VAD dominance: dominance
Pathos	Impact	gain_framing	Impact/enabling phrases	Lexicon IMPACT (binary): make a difference, have an impact, change my life
Pathos	Impact	gain_framing	Gain/benefit language	Lexicon IMPACT_GAIN (binary): benefit, benefits, advantage
Pathos	Impact	gain_framing	Soft gain frames	Regex RE_SOFT_GAIN (binary): better solution, better approach, better strategy
Pathos	Impact	loss_framing	Loss/cost language	Lexicon IMPACT_LOSS (binary): loss, waste, cost
Pathos	Impact	threat_severity	Threat/severity language	Lexicon IMPACT_THREAT (binary): catastrophic, deadly, fatal
Pathos	Impact	future_projection	Future-oriented modal cues	Lexicon IMPACT_WILL (binary): will, going to, gonna
Pathos	Impact	future_projection	Time-span projections	Regex RE_TIME_SPAN (binary): within 3 days; in two weeks; over 1 year
Pathos	Impact	future_projection	Future-frame imagery	Regex RE_FUTURE_FRAME (binary): future generations, imagine a world, too late
Pathos	Scarcity/Urgency	temporal_urgency	Urgency time cues	Lexicon URGENCY_TIME: urgent, urgently, immediately
Pathos	Scarcity/Urgency	temporal_urgency	Near-time cues	Lexicon SPEC_PSYCH_NEAR: minute, hour, day
Pathos	Scarcity/Urgency	exclusivity_quantity	Scarcity/exclusivity cues	Lexicon SCARCITY_QUANTITY: only, just, few left
Pathos	Engagement	identification	First-person identity cues	LIWC I: I
Pathos	Engagement	identification	Collective pronouns	Regex RE_WE: we, our, us
Pathos	Engagement	identification	Collective-care phrases	Regex RE_WE_care: we care about, our future, our children

Continued on next page

Table A2 – *continued from previous page*

Appeal	Dimension	Subfeature	Linguistic Cue	Operationalization
Pathos	Engagement	self_reference	Second-person reference	LIWC You: You
Pathos	Engagement	inquiry	Question marks	Regex RE_QUESTION (binary): ?
Pathos	Engagement	past	Past-tense language	LIWC Past: Past
Pathos	Engagement	imagery	Visual perception words	LIWC See: See
Pathos	Engagement	imagery	Imagery prompts	Regex RE_IMAGERY: imagine, picture, visualize
Pathos	Engagement	characters	Personal pronouns/characters	LIWC Ppron: Ppron
Pathos	Reciprocity	direct_promise	Direct reciprocity promise patterns	Regex RECIPROCITY_PATTERNS (binary): return the favor; do the same for you; pass it on
Pathos	Reciprocity	mutual_benefit	Mutual benefit/exchange cues	Lexicon REC_MUTUAL_EXCHANGE (binary): together, mutual, shared
Pathos	Reciprocity	concession_framing	Concession/compromise cues	Lexicon REC_CONCESSION (binary): least i can do, compromise, meet in the middle
Pathos	Reciprocity	concession_framing	Soft concession frames	Regex RE_SOFT_CONCESSION (binary): would you be open to, just wanted to, might be helpful
Pathos	Propaganda	emotional_charge	Loaded language	Lexicon PROP_LOADED (binary): corrupt, evil, traitor
Pathos	Propaganda	emotional_charge	Name-calling cues	Lexicon PROP_NAMECALL (binary): idiot, fool, liar
Pathos	Propaganda	emotional_charge	Fear/loss/threat cues	Lexicon PROP_FEAR (binary): threat, danger, destroy; Lexicon IMPACT_LOSS (binary): loss, waste, cost; Lexicon IMPACT_THREAT (binary): catastrophic, deadly, fatal
Pathos	Propaganda	logical_distortion	Whataboutism cues	Lexicon PROP_WHATABOUT (binary): what about, meanwhile, they ignore
Pathos	Propaganda	logical_distortion	Strawman cues	Lexicon PROP_STRAW (binary): they claim, they say, but actually
Pathos	Propaganda	logical_distortion	Vagueness-as-authority cues	Lexicon PROP_VAGUE (binary): some say, experts agree, it is known
Pathos	Propaganda	logical_distortion	Oversimplification cues	Lexicon PROP_OVERSIMPLIFY (binary): it's simple, the only reason, just because
Pathos	Propaganda	heuristic_identity_appeals	Flag/national identity cues	Lexicon PROP_FLAG: nation, country, people
Pathos	Propaganda	heuristic_identity_appeals	Bandwagon cues	Lexicon PROP_BANDWAGON: everyone, all of us, people know
Pathos	Propaganda	heuristic_identity_appeals	SIogan cues	Lexicon PROP_SLOGAN: make america great again, no more, never forget
Pathos	Propaganda	heuristic_identity_appeals	Propaganda authority cues	Lexicon PROP_AUTH: according to, scientists say, researchers found

Table A2: Persuasion Index (PI) methods table with one linguistic cue per row. **Appeal** is the top-level Aristotelian category (logos, ethos, pathos); **Dimension** is one of the 15 PI dimensions (e.g., EVIDENCE, ENGAGEMENT); **Subfeature** is one of the 55 finer-grained components that compose a dimension; **Linguistic Cue** names the textual phenomenon being measured (a lexical class, syntactic pattern, or punctuation marker); **Operationalization** gives the regex, lexicon lookup, or rule used to detect it. A subfeature score is the aggregate of all cue measurements assigned to that subfeature, normalized to per-100-token density, unless the cue is marked (binary), in which case the feature is computed as a presence/absence indicator (0/1) per document. Example words and phrases shown in the Operationalization column are drawn from our hand-crafted seed lexicons; the full lexicons (including LLM-expanded entries reviewed by domain experts) are released with the codebase.

B Datasets

We construct and evaluate the PI using four publicly released datasets (shown in Table A3) that together cover pairwise convincingness, topic- and side-informed arguments, and expert or crowd ratings of individual arguments.

UKPConvArg1 (UKP) (Habernal and Gurevych, 2016). UKPConvArg1 is a **pairwise** convincingness dataset (n=11,650) in which each instance contains two short arguments and crowd judgments indicating which argument is more convincing. Because arguments often appear in multiple pairwise comparisons within a topic, the dataset supports both topic-level and side-level analyses. UKPConvArg1 provides our primary supervision signal for estimating PI weights.

Table A4 reports the full topic distribution in UKPConvArg1, listing the original topic identifiers, their shortened display labels, and the number of unique arguments per topic. The display labels are used in all figures and coefficient plots throughout the paper for readability.

ChangeMyView (CMV) (Tan et al., 2016). Rather than the full CMV corpus, we use the ConvoKit "Winning Arguments" subset (n=8,526), where each reply is labeled according to whether the original poster awarded a Δ, signaling successful persuasion. Although CMV is not pairwise, the Δ annotation naturally produces a binary persuasive outcome: positive Δ indicates a persuasive (winning) argument, and vice versa.

IBM Argument Quality (IBM) (Gretz et al., 2020). The IBM Argument Quality dataset (n=14,003) contains expert-annotated assessments of argument quality and includes **pairwise** comparisons within topics. Although framed as argument quality rather than persuasion, prior work shows that these dimensions correlate strongly with theoretical constructs of persuasive effectiveness, making it suitable for evaluating PI's interpretability and generalization.

Table A5 reports the full topic distribution in IBM Argument Quality, listing the original topic identifiers, their shortened display labels, the stance either PRO or CON, and the number of unique arguments per topic. The display labels are used in all figures and coefficient plots throughout the paper for readability.

Anthropic Dataset (Anthropic) (Durmus et al., 2024). We include the publicly released Anthropic Persuasion dataset (n=3,882), a collection of human- and model-generated arguments annotated for persuasive intent and strategy. Each instance consists of a single argument paired with a persuasion label or tag, enabling argument-level analysis without requiring pairwise comparisons. Because the dataset directly encodes persuasive behaviors across a wide range of topics and styles, it provides a complementary evaluation setting for testing the generality and interpretability of the Persuasion Index.

C Persuasiveness Signal Construction

Our datasets provide two types of supervision: pairwise convincingness comparisons (UKP, IBM) and single-instance persuasion outcomes (CMV, Anthropic).

Pairwise datasets (UKP and IBM). For each argument pair (A,B), we construct a difference vector $\Delta\mathbf{x} = \mathbf{x}_A - \mathbf{x}_B$ from each argument's PI feature vectors, with binary label y=1 if argument A is judged more convincing and y=0 otherwise. This formulation encodes the contrastive signal between paired arguments without aggregating win–loss statistics across the full dataset, avoiding the confound where an argument's label depends on which other arguments it happens to be paired with.

CMV. Each reply is labeled y=1 if the original poster awarded a Δ, signaling that the reply successfully changed their view, and y=0 otherwise.

Anthropic. Each argument includes participant attitude ratings before and after exposure. We first remove participants with extreme baseline attitudes, those who initially selected the strongest agree or strongest disagree positions, as these show strong ceiling and floor effects unlikely to reflect genuine persuasion: 97% of participants with the strongest initial agreement maintained that position after reading, and 74.4% of those with the strongest initial disagreement did the same. After filtering, we label y=1 if the post-exposure attitude rating improves by more than 1 point relative to the pre-exposure baseline, capturing substantive attitude change while reducing noise from marginal shifts.

D Model

We evaluate the Persuasion Index by comparing logistic regression classifiers fit on PI feature vectors

Dataset	Size	Domain	Label Type	Persuasiveness Signal	Setting	Annotator
UKPConvArg1 (Habernal and Gurevych, 2016)	11,650 pairs	Web debate forums (16 topics)	Pairwise winner	Crowd convincingness judgment	Pairwise	Crowd
CMV (Tan et al., 2016)	8,526 replies	Reddit /r/CMV (diverse topics)	Binary (Δ / no Δ)	Original poster Δ award	Single-arg	Community
IBM Arg. Quality (Gretz et al., 2020)	14,003 args	Debating (71 topics)	Quality score + pairs	Expert quality rating	Pairwise	Expert
Anthropic Persuasion (Durmus et al., 2024)	3,882 args	Diverse (human & LLM-gen.)	Continuous (Δ rating)	Pre/post attitude shift	Single-arg	Human subjects

Table A3: Overview of the four evaluation datasets.

Original Identifier	Display Label	N
is-the-school-uniform-a-good-or-bad-idea-	School Uniforms	70
pro-choice-vs-pro-life	Abortion Debate	70
human-growth-and-development-should-parents-use-spanking-as-an-option-to-discipline-	Spanking Discipline	70
is-it-better-to-have-a-lousy-father-or-to-be-fatherless-	Father vs Fatherless	70
if-your-spouse-committed-murder-and-he-or-she-confided-in-you-would-you-turn-them-in-	Reporting a Spouse	70
gay-marriage-right-or-wrong	Gay Marriage	70
personal-pursuit-or-advancing-the-common-good-	Individual vs Common Good	70
william-farquhar-ought-to-be-honoured-as-the-rightful-founder-of-singapore	Farquhar as Founder	69
evolution-vs-creation	Evolution Debate	68
india-has-the-potential-to-lead-the-world-	India Leading	67
tv-is-better-than-books	TV vs Books	62
firefox-vs-internet-explorer	Firefox vs IE	62
christianity-or-atheism-	Christianity vs Atheism	62
ban-plastic-water-bottles	Ban Plastic Bottles	60
is-porn-wrong-	Porn Morality	56
should-physical-education-be-mandatory-in-schools-	Mandatory PE	56
Total		**1,052**

Table A4: Topic distribution in UKPConvArg1. N is the number of unique arguments per topic after deduplication across pairwise comparisons. Display labels are shortened versions of the original identifiers used in all figures and coefficient plots throughout the paper.

Original Source File	Topic	Stance	N
Social-media-brings-MORE-GOOD-than-harm	Social Media	PRO	654
Social-media-brings-MORE-HARM-than-good	Social Media	CON	520
We-should-support-information-privacy-laws-(PRO)	Information Privacy	PRO	253
We-should-promote-autonomous-cars-(CON)	Autonomous Cars	CON	205
We-should-adopt-vegetarianism-(PRO)	Vegetarianism	PRO	163
We-should-abandon-vegetarianism-(CON)	Vegetarianism	CON	146
We-should-legalize-doping-in-sport-(PRO)	Doping in Sport	PRO	141
Flu-vaccination-should-be-mandatory-(PRO)	Flu Vaccination	PRO	131
We-should-ban-doping-in-sport-(CON)	Doping in Sport	CON	129
We-should-adopt-cryptocurrency-(PRO)	Cryptocurrency	PRO	127
Flu-vaccination-should-not-be-mandatory-(CON)	Flu Vaccination	CON	121
We-should-abandon-cryptocurrency-(CON)	Cryptocurrency	CON	120
We-should-limit-autonomous-cars-(PRO)	Autonomous Cars	PRO	120
Gambling-should-be-banned-(PRO)	Gambling	PRO	109
We-should-ban-the-sale-of-violent-video-games-to-minors-(PRO)	Sale Violent Games	PRO	101
Gambling-should-not-be-banned-(CON)	Gambling	CON	94
We-should-allow-the-sale-of-violent-video-games-to-minors-(CON)	Sale Violent Games	CON	91
We-should-ban-fossil-fuels-(PRO)	Fossil Fuels	PRO	69
Online-shopping-brings-more-harm-than-good-(PRO)	Online Shopping	PRO	69
We-should-not-ban-fossil-fuels-(CON)	Fossil Fuels	CON	64
Online-shopping-brings-more-good-than-harm-(CON)	Online Shopping	CON	63
We-should-discourage-information-privacy-laws-(CON)	Information Privacy	CON	43
Total			**3,354**

Table A5: Source file distribution in the IBM Argument Quality dataset. Each row corresponds to one topic–stance combination. Topic labels are used in all figures and coefficient plots throughout the paper.

to fine-tuned `RoBERTa` and zero-shot `GPT-4o`. All models are trained with an 80/20 train–test split using random seed 42, with stratification on binary persuasion labels.

Logistic Regression. Our primary predictive models are ℓ_2-regularized logistic regressions trained on either the 15-dimensional PI category-mean vector (`PI-mean`) or the full 55-dimensional sub-feature vector (`PI-sub`). Given feature vector $\mathbf{x}$ for argument i, the model estimates $P(y_i = 1 \mid \mathbf{x}_i) = \sigma(\mathbf{w}^\top \mathbf{x}_i + b)$, where $y_i \in \{0, 1\}$ is the binary persuasion label (§C). We use a STANDARDSCALER followed by LOGISTICREGRESSION with the `liblinear` solver, `max_iter=2000`, and `class_weight="balanced"` to correct for class imbalance. We report accuracy, F1 score, and ROC–AUC on the held-out test set.

Pairwise setting (UKP and IBM). For pairwise datasets, we construct difference-vector features by subtracting two PI feature vectors: $\Delta\mathbf{x} = \mathbf{x}_A - \mathbf{x}_B$, and train a standard logistic regression on $\Delta\mathbf{x}$ with binary label $y = 1$ if argument A is judged more

convincing. This formulation isolates the specific PI dimensions that distinguish winning arguments from losing ones, rather than modeling each argument's absolute score. We evaluate via pairwise accuracy, defined as the proportion of correctly identified winners across all held-out pairs.

RoBERTa Neural Baseline. We compare PI to RoBERTa-base (Liu et al., 2019). Arguments are tokenized with a maximum sequence length of 128 using RobertaTokenizer, and we fine-tune RobertaForSequenceClassification with a two-class softmax head for binary persuasion prediction. We train for two epochs using AdamW with learning rate 5×10^{-5} and batch size 8, with all random seeds fixed to 42 for reproducibility. At evaluation time, predicted probabilities $p(y{=}1)$ are obtained via softmax over the output logits. Rather than applying the default threshold of 0.5, we select the optimal decision threshold using Youden's J statistic (Youden, 1950): $J = \arg\max(\text{tpr} - \text{fpr})$.

GPT-4o Zero-shot Baseline. We additionally evaluate GPT-4o as a zero-shot baseline to benchmark PI against frontier LLM judgment without any fine-tuning. For pairwise datasets (UKP and IBM), we prompt GPT-4o to directly compare two arguments and output which is more persuasive. For single-instance datasets (CMV and Anthropic), we prompt the model to produce a binary judgment (1 = persuasive, 0 = not persuasive) for each argument independently. Full prompts are as follows:

```
TASK:
Read the two texts below and decide which one
    is more persuasive.

OUTPUT:
a1 = the first text ("a1") is more persuasive
    than the second text ("a2")
a2 = the second text ("a2") is more persuasive
    than the first text ("a1")

OUTPUT FORMAT:
Return only one label: a1 or a2.
Do not output anything else.
```

```
TASK:
Read the text below and decide whether or not
    it is persuasive.

OUTPUT:
0 = not persuasive
1 = persuasive

OUTPUT FORMAT:
Return only one character: 0 or 1.
Do not output anything else.
```

E Prediction Results

Table A6 reports prediction accuracy, precision, recall, and F1 score across all datasets and models.

Table A6: Classification performance on the positive class (precision, recall, F1) and overall accuracy, by dataset and model.

Dataset	Model	P	R	Acc	F1
Anthropic	GPT4o	0.326	**0.984**	0.328	**0.490**
	PI-mean	0.350	0.553	0.516	0.429
	PI-sub	0.359	0.541	0.533	0.432
	RoBERTa	**0.432**	0.373	**0.633**	0.400
CMV	GPT4o	0.563	0.436	0.528	0.492
	PI-mean	0.582	0.561	0.560	0.571
	PI-sub	**0.611**	**0.567**	**0.585**	**0.588**
	RoBERTa	0.559	0.389	0.520	0.459
IBM	GPT4o	**0.702**	0.503	**0.650**	0.586
	PI-mean	0.565	0.594	0.575	0.579
	PI-sub	0.581	**0.599**	0.590	**0.590**
	RoBERTa	0.538	0.142	0.518	0.225
UKP	GPT4o	**0.895**	**0.794**	**0.855**	**0.842**
	PI-mean	0.677	0.691	0.690	0.684
	PI-sub	0.768	0.768	0.775	0.768
	RoBERTa	0.590	0.541	0.595	0.565

E.1 Model Comparisons

We report three statistical tests to characterize how PI models compare to RoBERTa and GPT-4o. **McNemar's test** examines whether two models make systematically different errors on the same test instances, with $p < 0.05$ indicating significantly different error patterns. **Bootstrap difference tests** estimate the sampling distribution of the accuracy gap via 1,000 resamples, reporting 95% CIs and two-sided p-values for the observed difference. **TOST (Two One-Sided Tests)** tests practical equivalence within a ± 0.05 accuracy margin; $p < 0.05$ confirms the two models are equivalent, while failure to reject is inconclusive.

On UKP (Table A7), PI-sub significantly outperforms PI-mean (+8.5 pp, p=.001). PI-sub also significantly beats RoBERTa by 18.0 pp (p<.001) but trails GPT-4o by 8.0 pp (p=.001). No pair is confirmed equivalent under TOST.

CMV (Table A8) is the only dataset where PI-sub leads outright. PI-sub significantly outperforms PI-mean (+2.5 pp, p=.032), GPT-4o (+5.7 pp, p<.001), and RoBERTa (+6.5 pp, p<.001), demonstrating stronger generalization on organic Reddit persuasion data than either fine-tuned RoBERTa or zero-shot GPT-4o. PI-mean is significantly above RoBERTa (+4.0 pp, p=.028) and trends above GPT-4o (+3.1 pp, p=.055, borderline).

Comparison	Acc(M1)	Acc(M2)	Δ	Bootstrap 95% CI	Boot p	McNemar p	TOST p
PI-sub vs PI-mean	0.775	0.690	**+0.085**	[0.035, 0.135]	**0.001**	**0.001**	0.869
PI-sub vs RoBERTa	0.775	0.595	**+0.180**	[0.122, 0.237]	**0.000**	**0.000**	1.000
PI-sub vs GPT-4o	0.775	0.855	-0.080	[-0.128, -0.030]	**0.001**	**0.002**	0.864
PI-mean vs RoBERTa	0.690	0.595	**+0.095**	[0.035, 0.158]	**0.002**	**0.004**	0.909
PI-mean vs GPT-4o	0.690	0.855	-0.165	[-0.220, -0.110]	**0.000**	**0.000**	1.000
RoBERTa vs GPT-4o	0.595	0.855	-0.260	[-0.318, -0.205]	**0.000**	**0.000**	1.000

Table A7: UKP pairwise setting. Bold Δ marks significant wins for the first-listed model; bold p-values indicate $p < 0.05$.

Comparison	Acc(M1)	Acc(M2)	Δ	Bootstrap 95% CI	Boot p	McNemar p	TOST p
PI-sub vs PI-mean	0.585	0.560	**+0.025**	[0.003, 0.049]	**0.032**	**0.040**	0.077
PI-sub vs GPT-4o	0.585	0.528	**+0.057**	[0.025, 0.088]	**0.000**	**0.001**	0.650
PI-sub vs RoBERTa	0.585	0.520	**+0.065**	[0.030, 0.099]	**0.000**	**0.000**	0.801
PI-mean vs GPT-4o	0.560	0.528	+0.031	[-0.001, 0.064]	0.055	0.059	0.144
PI-mean vs RoBERTa	0.560	0.520	**+0.040**	[0.004, 0.075]	**0.028**	**0.031**	0.273
GPT-4o vs RoBERTa	0.528	0.520	+0.008	[-0.026, 0.043]	0.662	0.678	**0.008**

Table A8: CMV setting. Bold Δ marks significant wins for the first-listed model; bold p-values indicate $p < 0.05$.

Comparison	Acc(M1)	Acc(M2)	Δ	Bootstrap 95% CI	Boot p	McNemar p	TOST p
PI-sub vs PI-mean	0.590	0.575	+0.015	[-0.037, 0.068]	0.569	0.634	0.158
PI-sub vs RoBERTa	0.590	0.517	**+0.072**	[0.003, 0.142]	**0.038**	**0.046**	0.740
PI-sub vs GPT-4o	0.590	0.650	-0.060	[-0.122, 0.003]	0.062	0.079	0.615
PI-mean vs RoBERTa	0.575	0.517	+0.058	[-0.013, 0.125]	0.104	0.123	0.585
PI-mean vs GPT-4o	0.575	0.650	-0.075	[-0.140, -0.007]	**0.026**	**0.033**	0.767
RoBERTa vs GPT-4o	0.517	0.650	-0.133	[-0.193, -0.072]	**0.000**	**0.000**	0.992

Table A9: IBM pairwise setting. Bold Δ marks significant wins for the first-listed model; bold p-values indicate $p < 0.05$.

Comparison	Acc(M1)	Acc(M2)	Δ	Bootstrap 95% CI	Boot p	McNemar p	TOST p
PI-sub vs PI-mean	0.533	0.516	+0.017	[-0.022, 0.055]	0.391	0.434	0.095
PI-sub vs RoBERTa	0.533	0.633	-0.100	[-0.143, -0.058]	**0.000**	**0.000**	0.979
PI-sub vs GPT-4o	0.533	0.328	**+0.205**	[0.158, 0.252]	**0.000**	**0.000**	1.000
PI-mean vs RoBERTa	0.516	0.633	-0.117	[-0.161, -0.073]	**0.000**	**0.000**	0.997
PI-mean vs GPT-4o	0.516	0.328	**+0.188**	[0.142, 0.234]	**0.000**	**0.000**	1.000
RoBERTa vs GPT-4o	0.633	0.328	**+0.305**	[0.251, 0.359]	**0.000**	**0.000**	1.000

Table A10: Anthropic setting. Bold Δ marks significant wins for the first-listed model; bold p-values indicate $p < 0.05$.

On IBM (Table A9), `GPT-4o` is the strongest model overall, significantly outperforming `PI-mean` (+7.5 pp, p=.026) and `RoBERTa` (+13.3 pp, p<.001). `PI-sub` closes much of this gap: the 6.0 pp difference from `GPT-4o` does not reach significance (boot p=.062, McNemar p=.079), while `PI-sub` significantly beats `RoBERTa` (+7.2 pp, p=.038). `PI-sub` and `PI-mean` remain indistinguishable from each other (+1.5 pp, p=.569). No pair is confirmed equivalent under TOST, so non-significant comparisons here may be underpowered rather than truly equivalent.

Anthropic (Table A10) presents a striking inversion of the UKP ordering. `RoBERTa` is the strongest model here (acc. 0.633), significantly outperforming `PI-sub` (+10.0 pp, p<.001) and `PI-mean` (+11.7 pp, p<.001). Both PI models substantially outperform `GPT-4o`: `PI-sub` beats `GPT-4o` by 20.5 pp and `PI-mean` by 18.8 pp (p<.001 across all tests). `PI-sub` and `PI-mean` are indistinguishable from each other (+1.7 pp, p=.391), suggesting that mean-level scores are sufficient on this task.

F Full regression coefficients

Tables A11– A14 show all logistic regression coefficients for UKP, CMV, IBM, and Anthropic.

F.1 Full Coefficients by Topic

Figure A1 shows topic-level coefficient patterns for all 16 UKP topics. Several cross-topic regularities are visible. LOGIC/COHESION and EVIDENCE are among the most consistently positive predictors, appearing with moderate to strong positive coefficients across the majority of topics, reinforcing their role as stable Logos-oriented signals. SENTIMENT is broadly positive but shows greater variance, peaking in *Porn Morality* (1.31), *School Uniforms* (1.12), and *Abortion Debate* (0.80), while near-zero or negative in others. ENGAGEMENT is predominantly negative across topics, most strongly in *TV vs Books* (-0.84) and *Father vs Fatherless* (-0.87), consistent with the cross-dataset pattern observed in Section 5. OPPONENT'S VIEW shows near-zero coefficients across all topics, reflecting its rarity as a rhetorical move in this corpus. Topic-specific patterns also emerge: *Reporting a Spouse* is uniquely driven by *Scarcity/Urgency* (1.18), *Abortion Debate* by POLITENESS (1.43), and *Individual vs Common Good* by SPECIFICITY (0.76), suggesting that the rhetorical demands of each topic activate distinct persuasion dimensions beyond the cross-topic baseline.

Figure A2 shows topic-level coefficient patterns for all 11 IBM topics. Unlike UKP, where EVIDENCE and LOGIC/COHESION dominate broadly, IBM topics show more heterogeneous persuasion signatures with no single dimension consistently predictive across topics. SENTIMENT shows the most variable pattern, strongly positive for *Sale of Violent Games* (0.53) and *Information Privacy* (0.46) but negative for *Cryptocurrency* (-0.33) and *Vegetarianism* (-0.33). *Fossil Fuels* is driven by ARGUMENTATION (0.54) with strongly negative COMMITMENT (-0.83) and POLITENESS (-0.72), suggesting that on contested scientific policy, explicit claim structure matters while conciliatory framing does not. *Doping in Sport* similarly relies on ARGUMENTATION (0.42) and SENTIMENT (0.47) but penalizes COMMITMENT (-0.57), consistent with a topic where bold claims outweigh personal pledges. *Information Privacy* shows a distinctive pattern with strong positive SENTIMENT

Table A11: Logistic regression: Pairwise Mean-feature Differences (UKP)

	Pairwise outcome
	PI-mean
Argumentation	0.147**
	(0.071)
Authority/Credibility	0.093
	(0.065)
Commitment	0.182***
	(0.060)
Engagement	-0.140**
	(0.070)
Evidence	0.491***
	(0.064)
Impact	0.254***
	(0.064)
Logic/Cohesion	0.487***
	(0.076)
Politeness	0.128*
	(0.067)
Propaganda	0.054
	(0.063)
Reciprocity	0.054
	(0.059)
Scarcity/Urgency	0.052
	(0.063)
Sentiment	0.260***
	(0.068)
Specificity	-0.311***
	(0.069)
Style	0.263***
	(0.069)
const	-0.075
	(0.057)
Observations	1600
Pseudo R^2	0.178
Note:	*p<0.1; **p<0.05; ***p<0.01

Table A12: Logistic regression: Mean-feature (CMV)

	Single outcome
	PI-mean
Argumentation	-0.053*
	(0.028)
Authority/Credibility	-0.006
	(0.032)
Commitment	0.028
	(0.027)
Engagement	-0.017
	(0.031)
Evidence	0.148***
	(0.032)
Impact	0.084***
	(0.032)
Logic/Cohesion	0.021
	(0.028)
Opponent's View	-0.024
	(0.026)
Politeness	0.089***
	(0.029)
Propaganda	0.035
	(0.030)
Reciprocity	0.012
	(0.027)
Scarcity/Urgency	0.019
	(0.027)
Sentiment	-0.029
	(0.028)
Specificity	-0.017
	(0.028)
Style	0.092***
	(0.033)
const	0.095***
	(0.025)
Observations	6484
Pseudo R^2	0.014
Note:	*p<0.1; **p<0.05; ***p<0.01

(0.46), RECIPROCITY (0.42), and AUTHORITY/-CREDIBILITY (0.35), reflecting appeals to personal rights and institutional trust. OPPONENT'S VIEW shows near-zero coefficients across all topics, consistent with the UKP pattern. Together, these results reinforce that persuasive strategies are highly topic-dependent, and demonstrate PI's capacity to surface fine-grained rhetorical variation in an interpretable and auditable way.

G Feature Correlations

Figure A3 shows correlations between the 15 PI category-mean scores on UKP. Two clear correlation clusters emerge, broadly corresponding to the Logos and Pathos groupings. Within Logos, EVIDENCE, LOGIC/COHESION, and ARGUMENTATION are positively correlated, reflecting that well-structured arguments tend to deploy factual grounding and explicit reasoning jointly. Within Pathos, SENTIMENT, RECIPROCITY, IMPACT, and ENGAGEMENT form a second positive cluster, consistent with affective and relational appeals co-occurring in persuasive text. Notably, the two clusters show weak to negligible cross-cluster correlations, suggesting that Logos- and Pathos-oriented rhetorical strategies are largely orthogonal in this corpus. OPPONENT'S VIEW is isolated with near-zero correlations to all other dimensions, reflecting its rarity as a rhetorical move.

Figure A4 reveals finer-grained structure at the sub-feature level. Within-dimension correlations are strong by construction for some dimensions (e.g., SENTIMENT sub-features cluster tightly), but several cross-dimension correlations are also visible. AUTHORITY/CREDIBILITY.SPEECH_POWER shows a notable negative correlation with ARGUMENTATION sub-features, suggesting that powerless language co-occurs with sparser argumentative structure. ENGAGEMENT sub-features

Table A13: Logistic regression: Pairwise Mean-feature Differences (IBM)

	Pairwise outcome
	PI-mean
Argumentation	0.017
	(0.085)
Authority/Credibility	0.015
	(0.058)
Commitment	-0.178***
	(0.061)
Engagement	-0.267***
	(0.058)
Evidence	0.135***
	(0.052)
Impact	-0.062
	(0.052)
Logic/Cohesion	0.052
	(0.083)
Politeness	0.029
	(0.061)
Propaganda	0.059
	(0.052)
Reciprocity	0.037
	(0.056)
Scarcity/Urgency	-0.096
	(0.071)
Sentiment	0.224***
	(0.053)
Specificity	0.022
	(0.074)
Style	0.041
	(0.057)
const	-0.032
	(0.051)
Observations	1600
Pseudo R^2	0.030
Note:	*p<0.1; **p<0.05; ***p<0.01

Table A14: Logistic regression: Mean-feature (Anthropic)

	Single outcome
	PI-mean
Argumentation	0.013
	(0.041)
Authority/Credibility	0.118*
	(0.061)
Commitment	-0.016
	(0.041)
Engagement	-0.060
	(0.047)
Evidence	-0.032
	(0.061)
Impact	0.038
	(0.043)
Logic/Cohesion	-0.002
	(0.041)
Opponent's View	0.935
	(5582.655)
Politeness	-0.044
	(0.039)
Propaganda	-0.122***
	(0.041)
Reciprocity	0.027
	(0.040)
Scarcity/Urgency	-0.020
	(0.044)
Sentiment	-0.033
	(0.042)
Specificity	0.002
	(0.044)
Style	-0.036
	(0.043)
const	-0.696
	(200.503)
Observations	3105
Pseudo R^2	0.009
Note:	*p<0.1; **p<0.05; ***p<0.01

(`self_reference`, `past`, `characters`) form a tight positive cluster, consistent with their shared reliance on personal pronoun and narrative language. The overall sparsity of strong off-diagonal correlations supports the discriminant validity of the PI sub-feature space — most sub-features capture distinct rhetorical signals.

H Robustness of Coefficient Interpretation

To assess whether PI coefficients can be interpreted as estimates of rhetorical importance rather than artifacts of correlated features, we conduct two robustness analyses on UKP: (1) multicollinearity diagnostics via Variance Inflation Factors (VIFs) across the 15 dimensions, and (2) leave-one-out ablations that quantify the marginal contribution of each dimension to predictive performance.

H.1 Multicollinearity Diagnostics

We compute VIFs by treating each dimension in turn as the outcome variable in an auxiliary regression against the remaining 14 dimensions, after standardizing the design matrix to mean zero and unit variance. By convention, VIF > 10 indicates problematic multicollinearity, while VIF > 5 warrants attention. Table A15 reports per-dimension VIFs on UKP.

The median VIF of 3.41 indicates that most dimensions carry largely independent variance, but three dimensions exceed the conventional threshold of 10: STYLE (27.4), SENTIMENT (22.6), and LOGIC/COHESION (14.1). These elevated values reflect interpretable structural properties of the framework rather than redundancy in the taxonomy:

- STYLE aggregates surface markers including `length`, which correlates with the rate-based

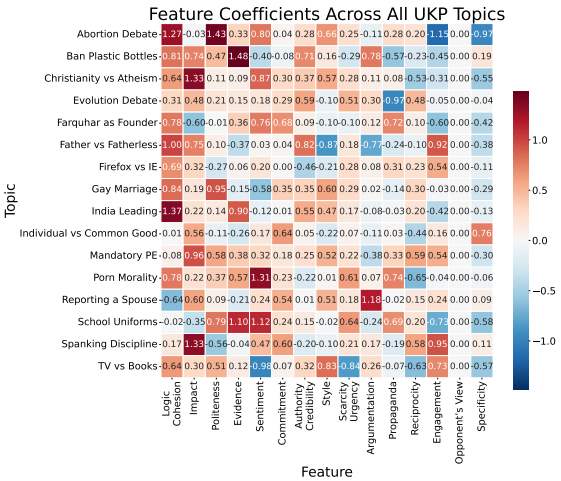

Figure A1: Coefficient Values of all features in UKP across different topics

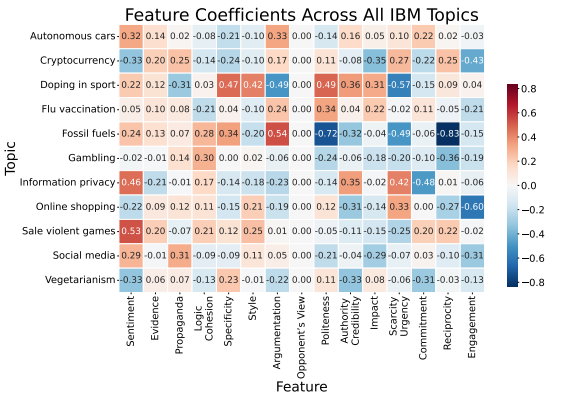

Figure A2: Coefficient Values of all features in IBM across different topics

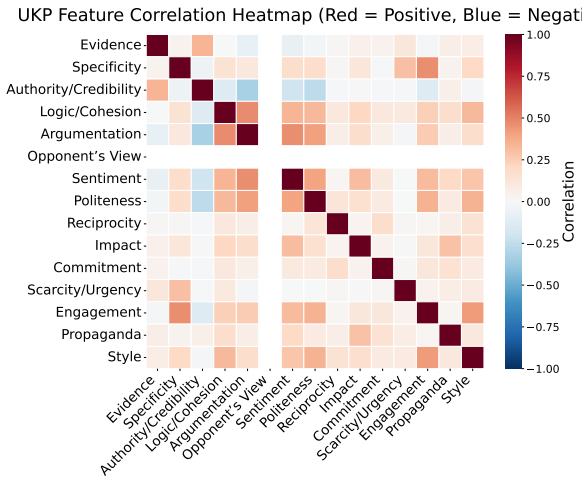

Figure A3: Pairwise Pearson correlations between the 15 PI category-mean scores on the UKP dataset.

Dimension	VIF
Style	27.39
Sentiment	22.64
Logic/Cohesion	14.09
Argumentation	8.36
Engagement	7.39
Specificity	4.61
Politeness	4.19
Evidence	2.64
Authority/Credibility	2.55
Propaganda	2.10
Scarcity/Urgency	1.96
Impact	1.89
Commitment	1.18
Reciprocity	1.17
Opponent's View	–
Median	3.41
Max	27.39

Table A15: Per-dimension VIFs on UKP. Opponent's View is omitted (NaN) because its sub-feature scores are near-constant on this dataset.

sub-features in every other dimension that longer arguments accumulate more matches across the framework as a whole.

- SENTIMENT aggregates nine partially overlapping sub-features spanning polarity (VADER), discrete emotions (anger, sadness, fear, joy), and dimensional affect (valence, arousal, dominance from NRC-VAD), with deliberate redundancy across these complementary affective scales.
- LOGIC/COHESION relies on connective-density sub-features that scale with text length, producing partial correlation with STYLE and other length-sensitive dimensions.

Two design choices mitigate the practical impact of these correlations on coefficient interpretation. First, our ℓ_2-regularized logistic regression explicitly shrinks correlated coefficients toward zero, stabilizing estimates in the regime where multicollinearity would otherwise produce instability. Second, the 95% Wald confidence intervals reported in Figures 3 and 5 explicitly reflect the standard-error inflation that VIF would predict; dimensions with high VIFs naturally exhibit wider CIs in our reported figures, allowing readers to calibrate confidence in individual estimates directly.

We therefore interpret PI coefficients at the *dimension* level as our primary analytic claim. Sub-feature-level interpretation, where some sub-features (e.g., `length` under STYLE, `named_entities` under EVIDENCE) may carry sur-

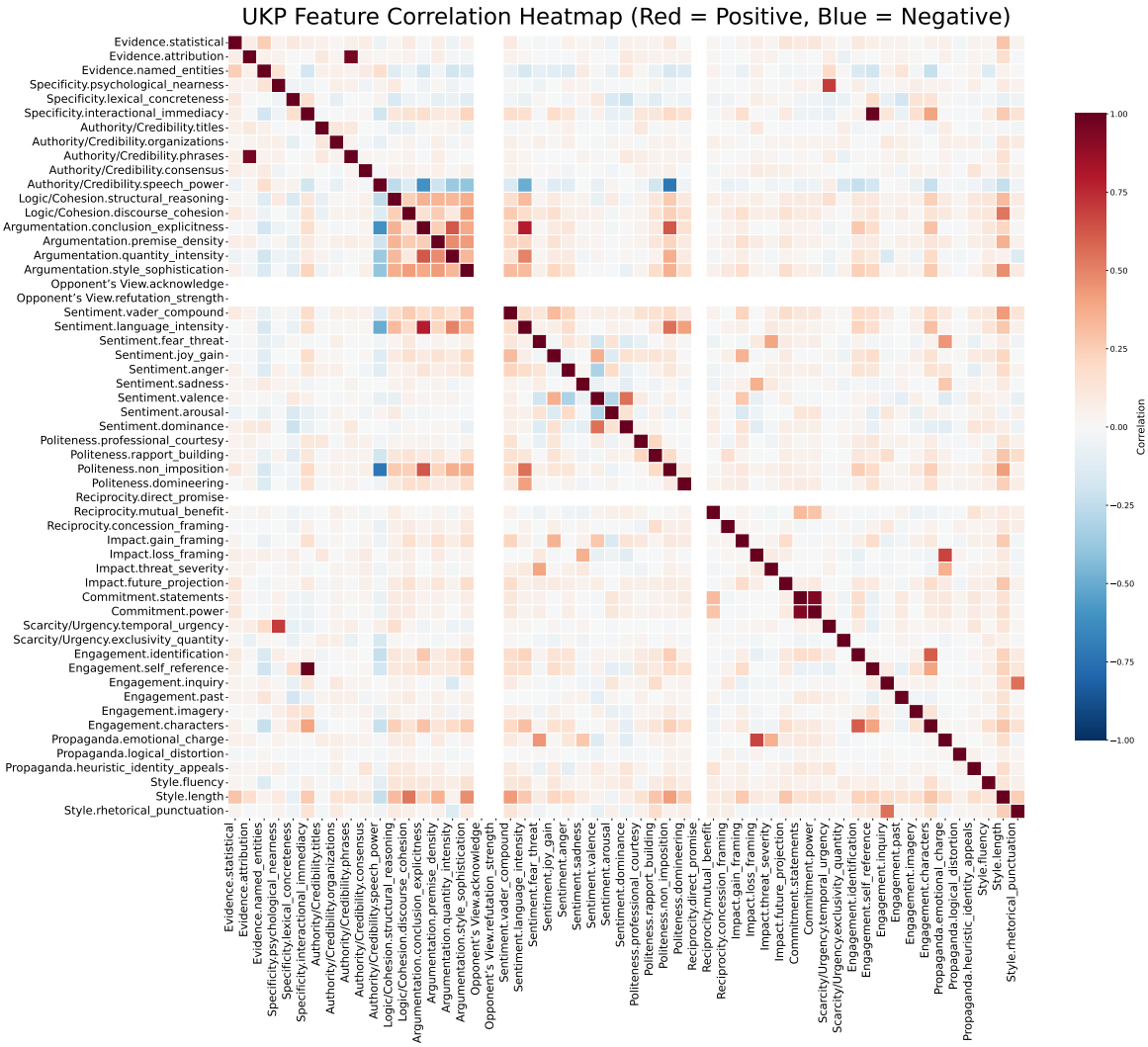

Figure A4: Pairwise Pearson correlations between the 55 PI sub-features on the UKP dataset.

face or topicality confounds in addition to rhetorical signal, requires the more targeted detectors that PI's modular design is intended to support (Section 4).

H.2 Leave-One-Out Dimension Ablations

To complement the multicollinearity diagnostics with a behavioral measure of dimension importance, we conduct leave-one-out ablations: for each of the 15 dimensions, we drop all of its constituent sub-features, refit the regression on the reduced feature matrix, and measure the resulting change in accuracy, F1, and AUC on UKP. Table A16 reports the results sorted by ΔAUC (most damaging first).

Ablation results converge with the coefficient analysis (Section 5). STYLE (–0.028 AUC) and SENTIMENT (–0.023 AUC) are the two most consequential dimensions for predictive performance on UKP, consistent with their large coefficient magnitudes and their elevated VIFs, where they carry both unique and shared variance with downstream persuasion judgments. ENGAGEMENT ranks third in importance (ΔAUC = –0.015) despite its consistently negative coefficient, suggesting that the dimension carries genuine discriminative signal even where its sign is interpretable as a negative predictor of persuasion in our corpora.

The remaining twelve dimensions each contribute marginally ($|\Delta\text{AUC}| \leq 0.007$), with three dimensions (SPECIFICITY, POLITENESS, ARGUMENTATION) showing slight improvements when removed, consistent with these dimensions carrying largely redundant signal already captured by the more central predictors. We retain them in the full taxonomy because PI is designed as a comprehensive theoretical framework rather than a minimal feature set; dimensions that contribute little to prediction on UKP may still carry interpretive value on

Removed Dimension	# Sub-features	Acc	F1	AUC	ΔAcc	ΔF1	ΔAUC
Style	3	0.7450	0.7358	0.8299	−0.0300	−0.0322	−0.0281
Sentiment	9	0.7375	0.7273	0.8352	−0.0375	−0.0407	−0.0228
Engagement	6	0.7750	0.7692	0.8426	0.0000	0.0012	−0.0154
Authority/Credibility	5	0.7725	0.7624	0.8538	−0.0025	−0.0056	−0.0042
Propaganda	3	0.7775	0.7700	0.8565	0.0025	0.0020	−0.0015
Reciprocity	3	0.7700	0.7617	0.8571	−0.0050	−0.0063	−0.0009
Impact	4	0.7700	0.7641	0.8577	−0.0050	−0.0039	−0.0003
Evidence	3	0.7625	0.7595	0.8578	−0.0125	−0.0085	−0.0002
Commitment	2	0.7700	0.7641	0.8580	−0.0050	−0.0039	0.0000
Opponent's View	2	0.7750	0.7680	0.8580	0.0000	0.0000	0.0000
Logic/Cohesion	2	0.7725	0.7649	0.8585	−0.0025	−0.0031	0.0005
Scarcity/Urgency	2	0.7750	0.7680	0.8588	0.0000	0.0000	0.0008
Argumentation	4	0.7750	0.7692	0.8598	0.0000	0.0012	0.0018
Politeness	4	0.7700	0.7641	0.8629	−0.0050	−0.0039	0.0049
Specificity	3	0.7850	0.7772	0.8651	0.0100	0.0092	0.0071
Full PI-sub	55	0.7750	0.7680	0.8580	—	—	—

Table A16: Leave-one-out ablations on UKP. Each row reports performance when the indicated dimension is removed and the regression is refit on the remaining feature matrix. Sorted by ΔAUC (most damaging removal first).

other domains (SENTIMENT's differential weighting across PRO and CON stances in Section 5) or under different audience conditions.

H.3 Interpretation and Caveats

Taken together, the VIF and ablation analyses support three claims about PI coefficient interpretation:

1. Coefficient estimates are stable enough at the dimension level to support comparative interpretation across datasets, topics, and stances, particularly for dimensions with VIF < 10, which covers twelve of fifteen dimensions.
2. For the three dimensions with elevated VIFs (STYLE, SENTIMENT, LOGIC/COHESION), individual coefficients should be read with the wider confidence intervals reported in the main figures, and conclusions should rely on the joint behavior of the dimension across multiple datasets rather than a single point estimate.
3. Sub-feature-level interpretation requires more caution than dimension-level interpretation. Sub-features such as `length`, `named_entities`, and frequency-based connective counts may capture verbosity or topicality confounds in addition to rhetorical signal. PI's modular design is explicitly intended to support upgrading these sub-features with confound-resistant detectors without altering the dimension-level taxonomy or its theoretical anchors.

Table A17: Inter-annotator agreement by LLM confidence level

Confidence	% Excl.	Krippendorff's α	N	All 3 Agree	Incl.	Excl.
High	5.4%	0.344	52	36	44	8
Low	67.6%	0.420	108	65	35	73
Med	31.0%	0.674	58	45	40	18

Bootstrap confidence intervals on PI coefficients, group-lasso analyses penalizing each Aristotelian triad as a unit, and sub-feature-level ablations are natural extensions of these checks, which we leave for future work.

I Lexicon Construction and Expansion

Here we provide full details for lexicon construction and validation.

Step 1: Manual Seed Lexicon Construction. We construct small seed lexicons based on foundational persuasion theories (Table A1). For example, the `Causal` sub-feature of the LOGIC AND COHESION dimension has a seed lexicon containing causal connectives ("because", "therefore") drawn from argumentation theory (Toulmin, 2003). Example seeds for all linguistic cues can be found in Table A2, and full seed and expanded lexicons are available in Supplementary Material.

Step 2: Manual Seed Validation. We conduct a three-stage validation of the seed lexicons to ensure construct validity. First, we review each seed lexicon by rating the fit of every term against its theoretical definition and removing terms with low face validity. Second, we run the seed-based PI on the datasets and examine the distribution and

frequency of each feature to flag implementation errors or incomplete lexicons. Third, we inspect sub-feature scoring at the top and bottom for each manipulated argument. We examine high-scoring arguments to confirm the detected signals reflect the intended persuasion dimension and inspect low-scoring arguments to identify missing patterns.

Step 3: LLM Expansion. Expansion begins with a structured prompt encoding the sub-feature label, theoretical anchor, and explicit inclusion and exclusion criteria derived from Steps 1 and 2 (example in Appendix I.1). Seven register- and morphology-conditioned prompts are issued to `GPT-5.4-mini` per lexicon category, targeting: formal single tokens, formal phrases, colloquial single tokens, colloquial phrases, domain-specific items (legal, academic, political), archaic or literary forms, and inflectional variants of canonical seeds. Since a single sub-feature may correspond to multiple lexicon categories (e.g., LOGIC/COHESION covers separate categories for causal, contrastive, and additive connectives), each category is expanded independently with its own structured prompt (Appendix I.2). Generated candidates are merged by lowercased surface forms and punctuation-based near-duplicates are collapsed. The model assigns a confidence level (high, medium, or low) to each candidate (prompt in Appendix I.3).

Step 4: Post-expansion Filtering. Three trained annotators (authors of this work) independently conducted a pilot annotation of 218 sampled items stratified across model-generated confidence tiers. The task involved binary labeling for whether or not a lexical item belongs to the sub-feature for which it was generated. All disagreements were resolved via consensus-coding. Overall Krippendorff's $\alpha = 0.558$ across three annotators, with majority agreement (at least 2 of 3) on all 218 items (67.0% unanimous) (Table A17). Rejection rates diverge substantially by confidence level: high-confidence items show a 5.4% rejection rate while low-confidence items reach 67.6%. Based on these rates, high-confidence items are admitted after a spot-check and low-confidence items are removed outright. For the 1,440 medium-confidence items, for which annotator agreement on the pilot sample was highest ($\alpha = 0.674$), remaining items were divided equally for independent annotation. Closed subfeature categories, such as "by the way", are dropped directly. The final lexicon takes the union of audited LLM-expanded candidates and the seed lexicon from Step 1.

Table A18: Sub-feature internal consistency (lexicon-driven features only), sorted by Pearson r (worst first). Computed on UKP ($n = 1{,}052$) using two random half-splits.

Feature	r	p	MAD
Scarcity/Urgency.exclusivity_quantity	−0.023	.460	0.157
Propaganda.logical_distortion	−0.002	.957	0.004
Impact.loss_framing	0.012	.700	0.073
Commitment.statements	0.027	.389	0.079
Reciprocity.concession_framing	0.038	.217	0.057
Argumentation.conclusion_explicitness	0.067	.029	0.369
Impact.threat_severity	0.079	.011	0.021
Authority/Credibility.phrases	0.089	.004	0.080
Impact.gain_framing	0.096	.002	0.220
Argumentation.premise_density	0.097	.002	0.452
Propaganda.emotional_charge	0.099	.001	0.137
Scarcity/Urgency.temporal_urgency	0.101	.001	0.155
Authority/Credibility.speech_power	0.103	<.001	0.402
Logic/Cohesion.structural_reasoning	0.135	<.001	0.413
Politeness.professional_courtesy	0.137	<.001	0.075
Propaganda.heuristic_identity_appeals	0.163	<.001	0.243
Politeness.rapport_building	0.165	<.001	0.264
Politeness.non_imposition	0.166	<.001	0.427
Sentiment.language_intensity	0.201	<.001	0.362
Specificity.psychological_nearness	0.205	<.001	0.226
Reciprocity.mutual_benefit	0.215	<.001	0.031
Argumentation.style_sophistication	0.228	<.001	0.217
Evidence.attribution	0.303	<.001	0.075
Commitment.power	0.305	<.001	0.038
Argumentation.quantity_intensity	0.314	<.001	0.288
Impact.future_projection	0.332	<.001	0.171
Evidence.statistical	0.343	<.001	0.286
Politeness.domineering	0.346	<.001	0.183
Logic/Cohesion.discourse_cohesion	0.838	<.001	0.069
Sentiment.fear_threat	0.844	<.001	0.021

Step 5: Final Lexicon Validation. A split-half internal consistency analysis verifies lexicon robustness. Each sub-category lexicon is partitioned into two equal, non-overlapping halves. The corpus is then scored independently using each half. Features driven entirely by non-lexical resources are excluded to prevent inflated results. The correlation between aggregate mean vectors is $r = 0.90$ ($p < 0.001$) across 30 lexicon-driven sub-features. Across the 12 mean-features, the correlation reaches $r = 0.97$ ($p < 0.001$). These confirm that the final lexicons preserve stable feature scales across the full PI framework. Sub-feature internal consistency is shown in Table A18

I.1 Category Definitions for LLM Expansion

Below is one category definition example to expand the seed lexicon "EVI_UNITS". Full category definitions are available in Supplementary Material.

```
"EVI_UNITS": {
"parent_appeal": "Logos",
"parent_dimension": "Evidence",
"parent_subfeature": "statistical",
"function": "Units of measurement, statistical
```

```
    markers, currencies, and quantitative
    indicators that ground claims in verifiable
    numerical evidence.",
"theory_anchor": "Zebregs et al. (2015) on
    statistical evidence; targets cognitive
    belief change via factual grounding.",
"include": [
  "SI / imperial units of measure (kg, mi, ml,
      ft, C, mph)",
  "Currency symbols and codes (USD, EUR, \$,
      JPY)",
  "Statistical indicators (p-value, percentage,
      ratio, coefficient, significance)",
  "Magnitude words used in numerical context
      (million, billion, thousand)"
],
"exclude": [
  "Polysemous letters with non-measurement
      meanings in dominant English use",
  "Generic quantifiers without measurement
      function (some, many, several) those
      belong to SPEC_VAGUE",
  "Number words used as ordinals/names (first,
      second as ranks)"
],
"common_drift_traps": [
  {
    "word": "size",
    "drift_type": "HYPERNYM_DILUTION",
    "why": "Generic noun, not a measurement unit"
  },
  {
    "word": "many",
    "drift_type": "TOPICAL_NEIGHBOR",
    "why": "Vague quantifier; belongs to
        SPEC_VAGUE"
  }
],
"canonical_seeds": [
  "%",
  "kg",
  "million",
  "percentage",
  "p-value",
  "USD"
]
}
```

I.2 Prompts for 7 slices

We broaden lexicon coverage by prompting with 7 distinct register "slices". Each slice is a configuration object substituted into the `[SLICE]` block of that template: `instructions` specifies the linguistic target (register, token granularity, and whether the slice generates positive examples or hard negatives), `type_field` and `register_hint` constrain the expected output, and `target_count` bounds the requested number of items. For each lexicon category we issue one LLM call per (category, slice) pair, then aggregate and deduplicate the returned items before human review.

```
{
"formal_single": {
    "instructions": (
        "Generate FORMAL-register SINGLE-WORD
            items. Targets: written, "
        "academic, legal, news-editorial,
            professional discourse. "
        "All items must be single tokens (no
            spaces). Set type='single'."
    ),
    "type_field": "single",
    "register_hint": "formal",
    "is_hard_negative": False,
    "target_count": 30,
},
"formal_phrase": {
    "instructions": (
        "Generate FORMAL-register N-GRAM items
            (multi-word expressions, "
        "fixed phrases, idiom-like locutions).
            Targets: 'in conclusion', "
        "'on the grounds that', 'it follows
            that'. Each item must contain "
        "at least one space (>=2 tokens). Set
            type='phrase'."
    ),
    "type_field": "phrase",
    "register_hint": "formal",
    "is_hard_negative": False,
    "target_count": 25,
},
"colloquial_single": {
    "instructions": (
        "Generate COLLOQUIAL / SPOKEN
            single-word items. Targets: spoken "
        "English, informal writing, social-media
            discourse. Include "
        "interjections, informal contractions
            where they function "
        "lexically (e.g., 'yeah', 'nah',
            'totally'). All items single "
        "tokens. Set type='single'."
    ),
    "type_field": "single",
    "register_hint": "colloquial",
    "is_hard_negative": False,
    "target_count": 20,
},
"colloquial_phrase": {
    "instructions": (
        "Generate COLLOQUIAL / SPOKEN N-grams
            that DIRECTLY perform "
        "the rhetorical function of THIS
            category, expressed in casual "
        "register. Each item >=2 tokens.\n\n"
        "CRITICAL: Do NOT output generic
            conversational fillers, "
        "hedges, or topic-management phrases
            (e.g. 'you know what', "
        "'I mean', 'to be honest', 'at the end
            of the day', 'the "
        "bottom line is', 'that being said')
            unless they LITERALLY "
        "encode the category's function. These
            items belong to other "
        "PI categories (Filler, By.The.Way,
            etc.) generating them "
        "here is category leakage.\n"
        "Test each candidate: 'If I delete the
            candidate from the "
        "sentence, does the {category} function
            disappear?' If the "
        "answer is no (the rhetorical function
```

```
            survives without it), "
        "do not output it.\n"
        "Examples of good output: for
            IMPACT_GAIN, 'pays off big', "
        "'comes out ahead', 'good deal'; for
            LOGIC_CAUSAL, 'that's "
        "why', 'that's because'.\n"
        "Set type='phrase'."
    ),
    "type_field": "phrase",
    "register_hint": "colloquial",
    "is_hard_negative": False,
    "target_count": 20,
},
"domain_specific": {
    "instructions": (
        "Generate items where THIS category's
            function appears in a "
        "specialised register (legal, academic,
            scientific, "
        "political, business). The item must
            still perform THIS "
        "category's rhetorical function not a
            different one that "
        "happens to share the domain.\n\n"
        "CRITICAL: Do NOT mechanically reach for
            examples from other "
        "PI categories (e.g., academic stance
            verbs like 'postulate' "
        "or 'we contend' are ARG_CLAIM, not
            LOGIC_CAUSAL). Re-read "
        "the FUNCTION block above and confirm
            the candidate satisfies "
        "it. Mark the source domain in the
            rationale (e.g., 'legal', "
        "'academic'). type may be 'single' or
            'phrase'."
    ),
    "type_field": "single_or_phrase",
    "register_hint": "domain",
    "is_hard_negative": False,
    "target_count": 20,
},
"archaic_literary": {
    "instructions": (
        "Generate items that perform the
            category function in a "
        "GENUINELY archaic, literary, or
            historically-marked register. "
        "Targets: items still encountered in
            religious, legal, or "
        "literary text but rare in modern
            conversational/journalistic "
        "prose.\n\n"
        "CRITICAL: Do NOT relabel modern,
            high-frequency items as "
        "archaic. 'therefore', 'hence', 'thus',
            'reward', 'prosperity', "
        "'advantage' are common modern words;
            they are not archaic and "
        "should NOT appear in this slice. True
            archaic examples: "
        "'wherefore', 'verily', 'forsooth',
            'whence', 'henceforth', "
        "'inasmuch as', 'thereupon', 'hereby'.
            If the category has no "
        "genuine archaic-register variants,
            output an EMPTY list "
        "this is the correct answer for many
            categories. Quality over "
        "count. type='single' or 'phrase'."
    ),
    "type_field": "single_or_phrase",
    "register_hint": "archaic",
    "is_hard_negative": False,
    "target_count": 15,
},
"inflectional": {
    "instructions": (
        "Generate INFLECTIONAL / DERIVATIONAL
            VARIANTS of the canonical "
        "seeds and other category members.
            Include third-person -s, "
        "past -ed, gerund -ing, agent -er/-or,
            negation prefixes when "
        "they preserve function.\n\n"
        "CRITICAL: If the canonical seeds are
            CLOSED-CLASS items "
        "(connectives, modals, demonstratives,
            fixed phrases) that do "
        "not inflect, output an EMPTY list.
            Examples of categories "
        "where you should output nothing:
            LOGIC_CAUSAL ('because' / "
        "'therefore' / 'hence' do not inflect);
            LOGIC_CONTRAST; "
        "LOGIC_REFERENCE; SPEC_PSYCH_NEAR
            (deictic 'now' / 'here'); "
        "PROP_BANDWAGON ('everyone' as
            quantifier).\n"
        "Never invent non-existent forms like
            'therefored' or "
        "'consequentlying'. Zero output is the
            correct answer when "
        "the category is non-inflectional.\n\n"
        "Only items whose function is genuinely
            preserved. Set "
        "type='inflection'."
    ),
    "type_field": "inflection",
    "register_hint": "neutral",
    "is_hard_negative": False,
    "target_count": 25,
},
```

I.3 Prompts to expand lexicons

The full prompt template, instantiated for the EVI_UNITS category and the formal_single slice, is shown below. Bracketed section headers in color demarcate the framework structure.

```
[CATEGORY]
Code: EVI_UNITS
Aristotelian appeal: Logos
PI dimension: Evidence
PI sub-feature: statistical
[FUNCTION]
Units of measurement, statistical markers,
    currencies, and quantitative indicators
    that ground claims in verifiable numerical
    evidence.
[THEORETICAL ANCHOR]
Zebregs et al. (2015) on statistical evidence;
    targets cognitive belief change via factual
    grounding.
[INCLUSION CRITERIA]
- SI / imperial units of measure (kg, mi, ml,
    ft, C, mph)
```

```
- Currency symbols and codes (USD, EUR, $, JPY)
- Statistical indicators (p-value, percentage,
    ratio, coefficient, significance)
- Magnitude words used in numerical context
    (million, billion, thousand)
[EXCLUSION CRITERIA]
- Polysemous letters with non-measurement
    meanings in dominant English use
- Generic quantifiers without measurement
    function (some, many, several) -- those
    belong to SPEC_VAGUE
- Number words used as ordinals/names (first,
    second as ranks)
[KNOWN DRIFT TRAPS]
- "size" (HYPERNYM_DILUTION): Generic noun, not
    a measurement unit
- "many" (TOPICAL_NEIGHBOR): Vague quantifier;
    belongs to SPEC_VAGUE
[CANONICAL SEEDS]
%, kg, million, percentage, p-value, USD
[SLICE: formal_single]
Generate FORMAL-register SINGLE-WORD items.
    Targets: written, academic, legal,
    news-editorial, professional discourse. All
    items must be single tokens (no spaces).
    Set type='single'.
Target output count: ~30 items (more is fine if
    quality holds; fewer is fine if the
    category is genuinely small).

[OUTPUT SCHEMA]
Return STRICT JSON in exactly this shape -- no
    prose, no fences:
{
  "items": [
    {
      "word": "<the lexicon item, lowercase
    unless capitalisation is intrinsic>",
      "register": "<formal | colloquial |
    domain | archaic | neutral>",
      "type": "<single | phrase | inflection |
    antonym | boundary>",
      "confidence": "<high | medium | low>",
      "rationale": "<one sentence; cite the
    specific INCLUDE or EXCLUDE rule that
    applies>"
    },
    ...
  ]
}
Field rules:
- "word" must be the surface form (not a regex
    pattern, not a list).
- "type" must equal "single" for this slice
    unless the slice is "domain_specific" or
    "archaic_literary" (in which case use
    "single" or "phrase").
- "register" should reflect the dominant
    register for the item.
- "confidence": use "high" only when no
    reasonable linguist would dispute
    admission; use "low" for borderline items
    that should go to human review.
- "rationale": one sentence. Cite the operative
    rule by paraphrase. No hedging.
Do not include items already given by the
    canonical seeds -- those are inputs, not
    outputs.
```

J Public Web Interface

Figure A5 and Figure A6 illustrate the public-facing Persuasion Index web interface, accessible at [URL not shown in this stage]. The tool allows users to input one or more arguments and receive real-time PI scores computed via server-side API calls. The dataset overview panel (Figure A5) provides dataset-level context for each of the four evaluation corpora, displaying the distribution of PI dimension scores across the corpus and surfacing the highest- and lowest-scoring dimensions as a reference baseline. Users can select a dataset context (e.g., UKPConvArg1) to calibrate their interpretation of scores against empirical norms. The analysis panel (Figure A6) supports pairwise or single-argument analysis. For each submitted text, the interface renders a radar chart across all 15 PI dimensions organized under Logos, Ethos, and Pathos, highlights the top-3 dimensions per text, and optionally computes a dataset-weighted persuasiveness score using the logistic regression coefficients estimated in our experiments. Raw scores can be exported as JSON for downstream use.

K Use of AI Assistants

AI Assistants are used for less than 10% of coding and writing in this work.

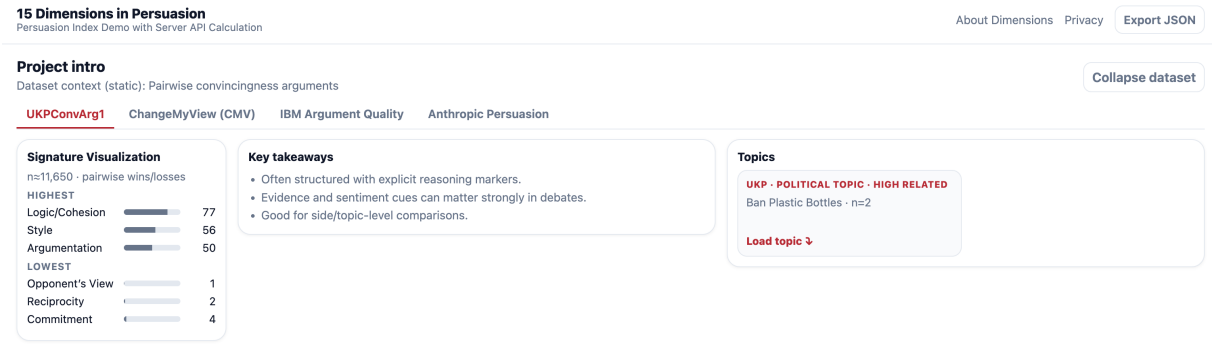

Figure A5: Dataset overview panel showing corpus-level PI dimension distributions and key takeaways.

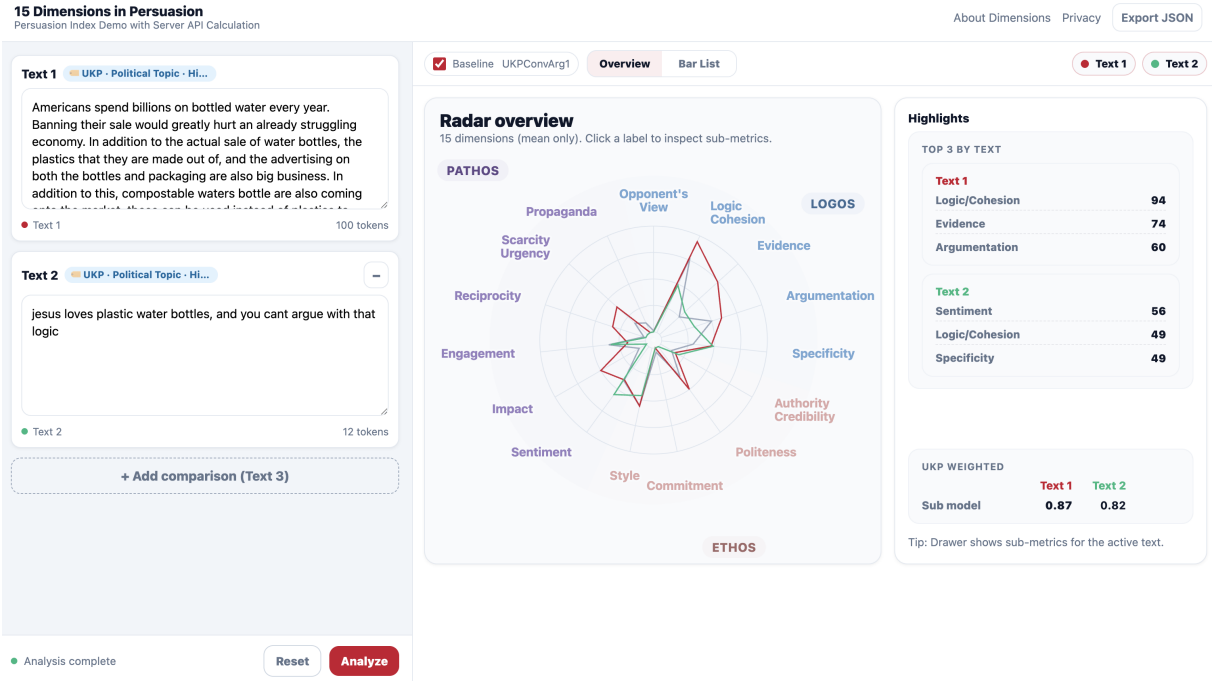

Figure A6: Analysis panel showing pairwise radar visualization across 15 PI dimensions with per-text highlights and weighted persuasiveness scores.