# You In or Out?

## Character Rhetoric and Audience Response in Attack on Titan*

**Peng-Ting Kuo**[1], **Gaeun Jung**[1], **and Hari Acharya**[2]

[1]Department of Government, University of Texas at Austin

[2]Department of Computer Science, University of Rochester

May 24, 2026

**Abstract**

This study examines how identity-based rhetoric employed by fictional characters shapes audience emotional responses in naturalistic online discourse. Using Attack on Titan as a case study, we construct an original dataset combining episode-level character dialogue with Reddit discussion threads and code rhetorical features — including threat framing, boundary marking, and solidarity appeals, using Gemini 2.5-pro. Employing a two-way fixed effects model with character and season fixed effects, we find that identity-based rhetorical intensity amplifies audience sentiment polarization and that solidarity appeals generate positive audience responses only when delivered by protagonist characters, consistent with parasocial identity reinforcement. Character-level fixed characteristics, however, account for the majority of variance in emotional responses, suggesting that narrative positioning conditions audience reception more strongly than episodic rhetorical content. Our findings contribute to the literature on narrative persuasion and Social Identity Theory, with implications for how fictional storytelling shapes audience moral alignment with political violence.

**Keywords**: LLM, factor analysis, computational social science, Attack on Titan

*Source code for reproducing all experiments is released at https://github.com/deankuo/Attack_on_Titan

# 1 Introduction

Political violence rarely emerges from purely rational calculation. A substantial body of research in social psychology suggests that group identity plays a foundational role in shaping how individuals perceive, justify, and support the use of violence against outgroups. Social Identity Theory, originally developed by Tajfel and John (2004), holds that individuals derive a significant portion of their self-concept from membership in social groups, and that this identification motivates behavior oriented toward maintaining or enhancing the perceived status and security of the ingroup. When in-group identity is threatened, individuals become more likely to endorse extreme measures in its defense, including the use of violence against those constructed as out-group enemies.

This foundational insight has been extended in several important directions. Research in social psychology suggests that conditions of uncertainty and perceived intergroup threat intensify ingroup identification and increase receptivity to narratives that frame violence as necessary for collective survival (Hogg 2007). Such dynamics are particularly powerful when out-groups are portrayed as existential threats, as individuals become more willing to justify extreme actions in defense of the in-group. Related research on identity fusion further shows that when personal and group identities become deeply intertwined, individuals exhibit greater support for extreme pro-group behavior, including violence (Horgan 2008). At the same time, Bandura's Bandura (2002) theory of moral disengagement explains how harmful actions toward out-groups become psychologically justifiable through mechanisms such as dehumanization, moral justification, and inevitability framing. Together, these perspectives suggest that narratives emphasizing ingroup survival, existential threat, and moral necessity may increase audience sympathy toward otherwise morally controversial violence.

The question of whether these mechanisms operate within fictional narratives, and whether exposure to such narratives shapes real-world moral alignment, remains underexplored. Media narratives have long been recognized as vehicles for shaping public attitudes and beliefs, yet existing research focuses on news media and political communication rather than fictional storytelling. This gap is consequential, because fictional narratives may in some respects constitute a more powerful persuasive environment than direct political communication. The fictional frame, rather than insulating audiences from persuasive influence, may paradoxically amplify it by reducing the counter-arguing that direct political rhetoric typically provokes.

This study investigates these dynamics through the lens of Attack on Titan (AoT), a Japanese anime series widely recognized for its morally controversial portrayal of war, genocide, and political violence, which generated sharp audience divisions over the justifiability of the protagonist's extreme actions. We argue that this division is not incidental but is systematically produced by the rhetorical strategies embedded in character dialogue. When characters frame violence through ingroup survival appeals, dehumanization, and fatalistic inevitability framing, audiences become more likely to morally align with extreme actions even when those actions would otherwise be condemned. We operationalize this moral alignment as the primary outcome of interest, measured through the proportion of audience comments expressing agreement with a character's justification of violence, and complement it with secondary measures of emotional intensity and audience polarization.

To test this, we construct an original dataset combining episode-level character dialogue with Reddit discussion threads and employ a two-way fixed effects (TWFE) model with character and episode fixed effects across all seasons of the series. Our findings reveal that the affective impact of rhetorical strategies is not uniform but is contingent on narrative character position. While identity-based rhetoric broadly intensifies the polarization of audience sentiment, consistent with Social Identity Theory's prediction that group boundary construction amplifies divergent evaluations, the emotional valence of that polarization depends on who delivers the rhetoric.

Solidarity appeals generate positive audience responses only when delivered by protagonist characters, suggesting that parasocial identification conditions rhetorical reception in ways that cannot be reduced to message content alone. Beyond its contribution to the study of narrative persuasion, this research carries broader implications for understanding how fictional storytelling, and by extension political narratives and propaganda, can shift real-world attitudes toward the legitimacy of violence.

We derive five testable hypotheses from Social Identity Theory and Narrative Persuasion Theory. The first set of hypotheses addresses whether identity-based rhetoric employed by fictional characters shapes audience emotional responses in online discussion spaces.

**Hypothesis 1 (H1a)** *Characters employing stronger identity-based rhetoric generate higher levels of negative audience affect, specifically anger and disgust, within Reddit discussion threads.*

**Hypothesis 2 (H1b)** *Characters employing stronger ingroup/outgroup rhetoric generate greater emotional polarization within audience discussions.*

**Hypothesis 3 (H1c)** *Characters employing stronger threat framing generate higher levels of fear expression within Reddit discussion threads.*

The second set of hypotheses examines whether the persuasive effects of rhetorical strategies are conditioned by the narrative position of the character delivering them. Narrative Persuasion Theory suggests that parasocial identification with protagonist characters shapes how audiences receive and evaluate rhetorical appeals, such that identical rhetoric may produce systematically different emotional responses depending on whether it is delivered by a protagonist or an antagonist.

**Hypothesis 4 (H2a)** *Threat framing delivered by protagonist characters generates weaker negative emotional responses than equivalent rhetoric delivered by antagonist characters.*

**Hypothesis 5 (H2b)** *Solidarity appeals delivered by protagonist characters generate higher levels of positive audience sentiment than equivalent rhetoric delivered by antagonist characters.*

## 2 Data Collection

To examine the rhetorical strategies employed by characters in AoT and their effects on audience responses, we construct a systematic data generation pipeline that enables reproducible collection and annotation of both narrative and audience data.

### 2.1 Transcript Data of AoT

We scraped the transcripts of all episodes from Season 1 to Season 4 of *Attack on Titan* from *Springfield!Springfield!*,[1] yielding a dataset of 15,000+ sentences across 89 episodes. These transcripts are English-language versions derived from the official Japanese subtitles, providing a consistent linguistic basis for downstream annotation. Figure 1 shows the transcript line length distribution.

1. Website: https://www.springfieldspringfield.co.uk. This is a website with archived transcripts of TV series and films.

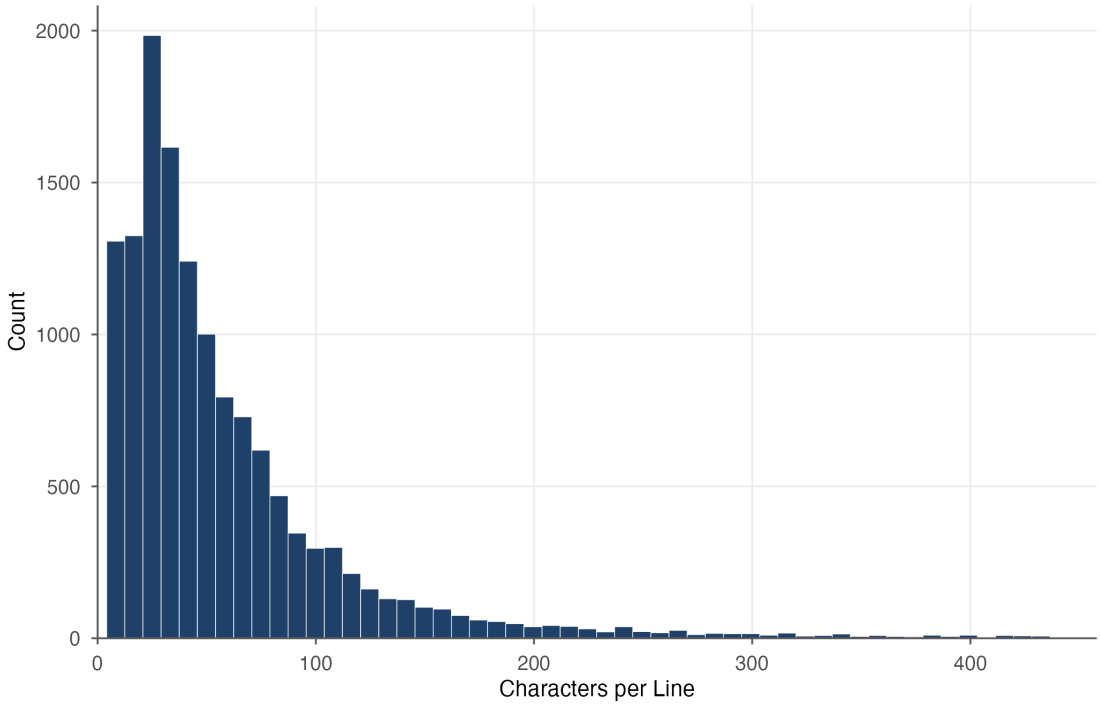



Figure 1: Distribution of Transcript Line Lengths

The raw transcripts do not contain speaker labels, as the source provides dialogue in plain text format. To attribute each utterance to its speaker, we employ Gemini-2.5-pro (Comanici et al. 2025) as an automated speaker identification model, supplying the full episode as context to provide sufficient narrative information for disambiguation. We then define a set of focal characters including: Eren Yeager, Hange Zoë, Armin Arlert, Jean Kirstein, Erwin Smith, Levi Ackerman, Reiner Braun, Theo Magath, Ymir, and Zeke Yeager, which selected on the basis that each sustains a substantive ideological arc across Seasons 1–4 and contributes meaningfully to the rhetorical dynamics of the narrative. Utterances attributed to characters outside this focal set, including minor characters and background speakers, are excluded from subsequent analysis.[2] Figure 2 shows the distribution of transcript lines after filtering across all episodes.

2. We identified a total of 78 unique characters across all episodes and retained the top 20 by dialogue volume for analysis.

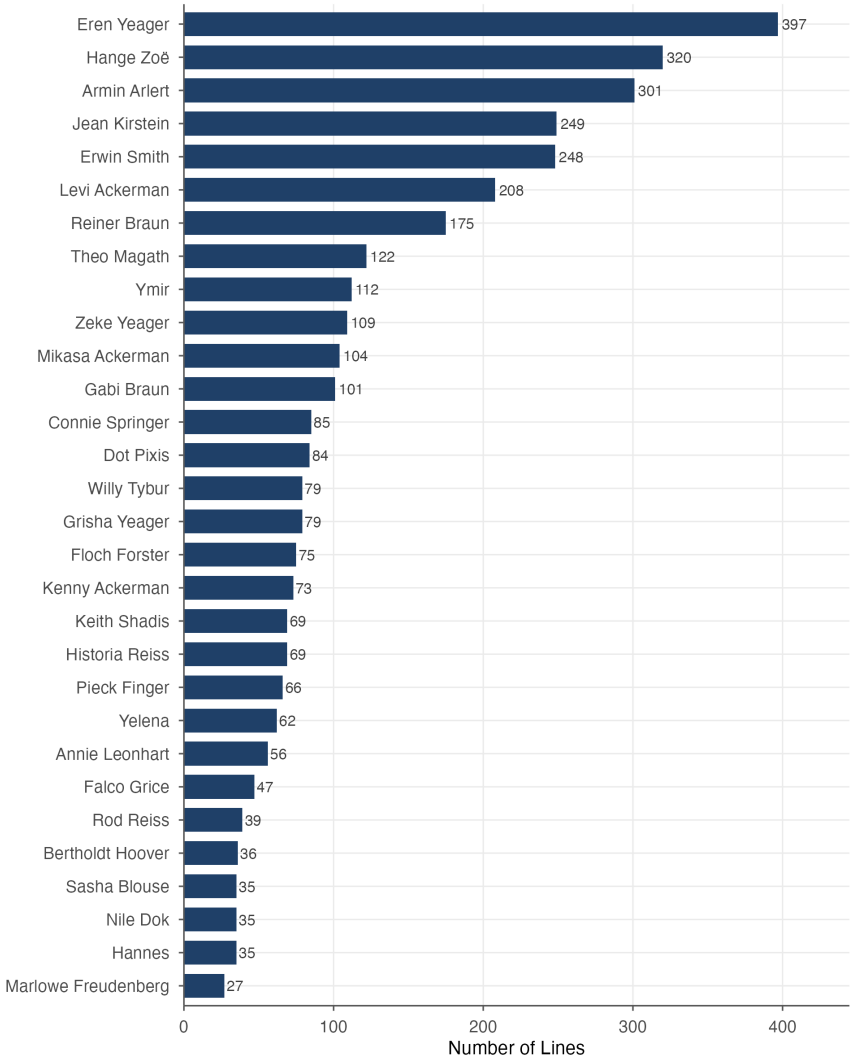



Figure 2: Top 30 Characters from AoT by Dialogue Lines

Following speaker identification, we apply two filtering criteria. First, we retain only sentences with a minimum length of 60 characters, ensuring sufficient semantic content for reliable rhetorical annotation; utterances below this threshold lack the syntactic and contextual information necessary for LLM-based scoring to be stable. Second, we retain only top 20 characters whose total qualifying sentence count across all episodes exceeds 65 sentences, guarding against unreliable cell-level estimates in the panel analysis. The resulting panel dataset consists of 794 observations at the episode $\times$ character level across all episodes.

## 2.2 Comment Data from Reddit

To capture audience reception of each episode, we construct a parallel corpus of Reddit discussions sourced from `r/attackontitan`, the largest English-language fan community devoted to the series. Reddit is well suited to this purpose because it organizes conversation into persistent, threaded structures indexed at the post and comment level, enabling systematic recovery of audience reaction at the resolution required for episode-level analysis. We retrieve both submissions and comments via the PullPush API[3], which provides the full historical archive without the listing-cap restrictions imposed by the official Reddit API. This is essential for our design, as high-engagement episode threads routinely exceed that cap by an order of magnitude.

For each of the 89 episodes in Seasons 1–4, we define a seven-day post-airdate window beginning at the recorded airdate timestamp and extending until either seven calendar days have elapsed or the subsequent episode airs, whichever occurs first. The seven-day length corresponds to the modal inter-episode interval of the series, with 75 of 88 episode-to-episode transitions occurring on a strict weekly cadence, while the next-episode cap prevents commentary on subsequent episodes from contaminating the focal window. For the season finale, where no subsequent airdate exists to bound the window, we apply the same seven-day length for comparability with the rest of the panel. Within each window we retrieve all submissions to `r/attackontitan` together with their full comment trees, and tag every record with its season, episode number, cumulative episode index, and the elapsed time since the airdate, enabling unambiguous joins to the dialogue panel constructed in Section 2.1.

Three structural edge cases warrant explicit treatment. First, Season 4 Episode 14 (S04E14) ("Savagery") aired on the same calendar day as Season 4 Episode 15 (S04E15), producing a window of zero duration under the strict cap rule; we treat this episode as missing for the Reddit panel rather than redefine the window asymmetrically. Second, Season 1 Episode 1 (S01E01) and Season 1 Episode 2 (S01E02), which aired in April 2013 when `r/attackontitan` had only several hundred subscribers, returned no archived comments within their seven-day windows; this absence reflects pre-existing subreddit sparsity rather than retrieval failure.

3. Website: https://pullpush.io. PullPush is the community-maintained successor to the discontinued Pushshift archive and exposes Reddit's historical record without the listing-cap restrictions of the official Reddit API, which truncates queries at approximately one thousand items.

Third, deleted and removed comments are retained in the corpus with their `author` field set to `[deleted]` for accurate volume counts but are excluded from any analysis that depends on comment content.

After retrieval, the raw corpus consists of 428,913 comment records authored by 63,115 unique users, distributed across 52,317 unique submissions and 85 episodes, spanning April 2013 through November 2023. The distribution of comments per episode is heavily right-skewed (median = 778, mean = 5,046, SD = 7,459), with approximately 91% of all retrieved comments concentrated in Seasons 3 and 4, consistent with the subreddit's growth trajectory and the rising cultural profile of the series during its final arcs.

To produce an analysis-ready corpus we apply four sequential preprocessing filters, with attrition tracked at each stage to support reproducibility. First, deleted and removed comments: those whose body has been replaced by `[deleted]` or `[removed]` and those authored by the placeholder `[deleted]` account are dropped, removing 12.6% of the raw corpus. Second, bot accounts are removed by combining an explicit blocklist of known Reddit bots (most prominently `AutoModerator`, which alone contributes 36,832 comments) with a username regex matching common bot suffixes; this step removes a further 9.0%. Third, comments with body length below 60 characters are dropped, mirroring the same 60-character threshold applied to transcript lines in Section 2.1 and ensuring that sentence-level sentiment and emotion scores are computed on units of comparable semantic density across both corpora; this is the largest single filter, removing 33.4% of the raw corpus. Fourth, a low-activity author filter retains only authors who contribute at least two surviving comments to the corpus, removing a further 19,704 records authored by single-use accounts that are disproportionately throwaways and unregistered visitors. After all four filters, 173,359 comments remain (40.4% of the raw corpus).

For alignment with the character-level dialogue panel, surviving comments are linked to the focal characters defined in Section 2.1 via a curated alias dictionary that maps each character to a set of first names, surnames, and established show-internal aliases (e.g., *Krista* for Historia, *Hanji* for Hange). Surnames that are ambiguous across multiple focal characters — *Ackerman*, *Yeager*, *Braun*, *Reiss* — are intentionally excluded from the alias set to avoid double-counting; a comment is tagged with every focal character it unambiguously mentions, allowing a single comment to contribute to multiple character cells when it discusses several characters. Of the 173,359 filtered comments, 66,718 (38.5%) reference at least one focal character.

Sentence-level emotion and sentiment scores from the pretrained classifiers described in Section 3 are then aggregated to the episode × character cell by taking the cell mean of each Ekman emotion probability and each sentiment class probability; in addition, we construct a net-valence score $\text{net}_{ij} = p(\text{positive})_{ij} - p(\text{negative})_{ij}$ at the comment level and report two cell-level summaries: the mean net valence and its within-cell standard deviation, the latter operationalizing the *audience polarization* outcome used in Hypothesis 1b. The resulting Reddit-side panel comprises 1,112 non-empty (episode × character) cells across the 20 focal characters, which we join to the speaker-side panel on the same keys to form the regression-ready dataset.

# 3 Measurement

## 3.1 Rhetorical Annotation

We annotate each qualifying sentence using both a commercial model, `GPT-5-mini` (OpenAI et al. 2024), and an open-source model, `Qwen2.5-32B-Instruct` (Hui et al. 2024; Yang et al. 2025), across two theoretical dimensions: moral disengagement mechanisms and in-group/out-group framing. `GPT-5-mini` represents the most cost-performance balanced option among closed-source models, while `Qwen2.5-32B-Instruct` serves as a strong open-source alternative with comparable performance. Each dimension is operationalized as a set of sub-indicators scored on a Likart scale (1-5), described in detail below.

For the open-source rhetorical annotation model, `Qwen2.5-32B-Instruct` was deployed on Texas Advanced Computing Center (TACC) across two NVIDIA A100 40GB GPUs using tensor parallelism, as the model's parameter count exceeds the memory capacity of a single device. Annotation was conducted in batches with a maximum output length of 8192 tokens per sentence-level prompt, including the full CoT instruction, few-shot exemplars, and the target sentence with a fallback output length up to 16000 tokens.

To enhance annotation reliability, we employ a structured prompting strategy that incorporates chain-of-thought (CoT) reasoning (Wei et al. 2022), step-by-step thinking (Kojima et al. 2022), and three-shot in-context learning (Brown et al. 2020; Gao, Fisch, and Chen 2021; Zhao et al. 2021). Recent research further demonstrates that persona injection improves LLM performance on downstream social science annotation tasks (Argyle et al. 2023; Gilardi, Alizadeh, and Kubli 2023; Törnberg 2025), and we incorporate this

technique accordingly. For each annotation dimension, the prompt instructs the model to first identify whether the relevant rhetorical feature is present, reason through the evidence for its presence or absence, and then assign a 1-5 label. This CoT structure encourages the model to surface its reasoning before committing to a label, reducing reflexive or context-insensitive responses (Wei et al. 2022). Each prompt additionally includes three few-shot exemplars per indicator: one clearly positive instance, one clearly negative instance, and one ambiguous intermediate case, drawn from a held-out set of AoT utterances annotated by the authors prior to the main annotation run, ensuring domain consistency. The full prompts for the annotation can be found in Appendix A.

For sentiment annotation, we apply two pre-trained models at the sentence level to both the transcript dialogue and the Reddit comment corpora, enabling parallel measurement of emotional tone in character speech and audience response.

For emotion classification, we use `j-hartmann/emotion-english-distilroberta-base` (Hartmann 2022), a model built on the DistilRoBERTa architecture (Liu et al. 2019) and trained on six diverse English-language datasets spanning multiple domains and text types. The model predicts Ekman's six basic emotions: anger, disgust, fear, joy, sadness, and surprise, plus a neutral class, outputting a calibrated probability score for each label rather than a hard classification. We retain the full probability vector for each sentence, which allows us to use continuous emotion intensity scores rather than collapsed categorical labels in subsequent analysis. For valence analysis, we use `cardiffnlp/twitter-roberta-base-sentiment-latest` (Card et al. 2022; Loureiro et al. 2022), also a RoBERTa-base model (Liu et al. 2019) trained on 124 million English tweets spanning January 2018 to December 2021 and fine-tuned for three-class sentiment classification (negative, neutral, positive). This model likewise outputs a probability distribution over the three classes, from which we derive a continuous net valence score defined as $p(\text{positive}) - p(\text{negative})$ for each sentence. Both models are applied to the transcript dialogue corpus and the Reddit comment corpus independently, yielding sentence-level emotion and valence scores for each, which are subsequently aggregated to the episode $\times$ character level as described below. Inference for both models was conducted on a single NVIDIA A100 40GB GPU, with sentences processed in batches of 32 to optimize throughput.

## 3.2 Index Construction via Factor Analysis

To assess the measurement structure of our annotation indicators and construct composite indices, we conduct exploratory factor analysis (EFA) (James et al. 2021) separately for each theoretical dimension at the sentence level, where sample size is sufficient for stable factor recovery. We use maximum likelihood estimation with oblique rotation (oblimin) to allow factors to correlate, consistent with theoretical expectations that moral disengagement mechanisms are empirically related despite their conceptual distinctness (Osborne 2015).

We retain a two-factor solution on the basis of both theoretical considerations and analytic parsimony. As shown in Figure 3, the scree plot indicates that only the first factor exceeds the parallel analysis retention threshold; however, we retain two factors on theoretical grounds consistent with prior empirical work (Caprara et al. 2009). Table 1 presents the factor loadings after oblimin rotation, with loadings below 0.30 suppressed.

Table 1: Moral Disengagement Factor Loadings

| **Indicator** | **Factor 1** | **Factor 2** | $h^2$ |
|---|---|---|---|
| Moral Justification | | 0.42 | 0.28 |
| Euphemistic Labeling | | | 0.15 |
| Advantageous Comparison | | | 0.03 |
| Displacement of Responsibility | | 0.33 | 0.09 |
| Diffusion of Responsibility | | 0.41 | 0.18 |
| Disregard of Consequences | 0.87 | | 0.75 |
| Dehumanization | | 0.34 | 0.16 |
| Attribution of Blame | | 0.49 | 0.21 |

The two factors exhibit a theoretically interpretable structure. As Table 1 shows, factor 1 is dominated by a single indicator: Disregard of Consequences, and captures the minimization or denial of the harmful outcomes of one's actions. Factor 2 is more broadly constituted, with meaningful loadings on Attribution of Blame ($\lambda = 0.49$), Moral Justification ($\lambda = 0.42$), Diffusion of Responsibility ($\lambda = 0.41$), Dehumanization ($\lambda = 0.34$), and Displacement of Responsibility ($\lambda = 0.33$); this factor captures a cluster of mechanisms oriented toward redistributing moral accountability: either by invoking higher purposes, spreading responsibility across collectives, or directing blame toward victims and out-groups. Two indicators, Euphemistic Labeling and Advantageous Comparison, exhibit low communality and do not load meaningfully on either

factor, suggesting that these mechanisms are either infrequently deployed in the AoT narrative or are difficult to reliably distinguish at the sentence level through LLM-based annotation. We retain these indicators in the annotation scheme for theoretical completeness but note that their contribution to the composite index is negligible.

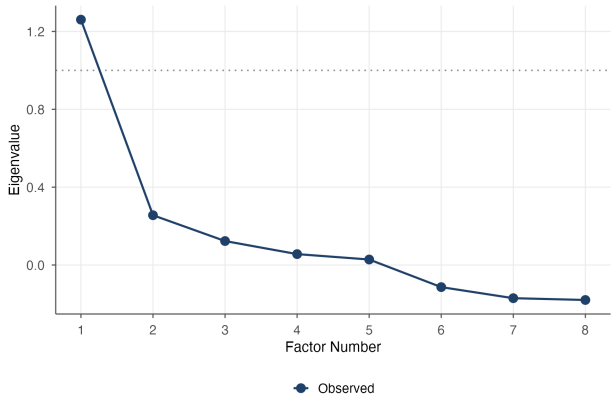
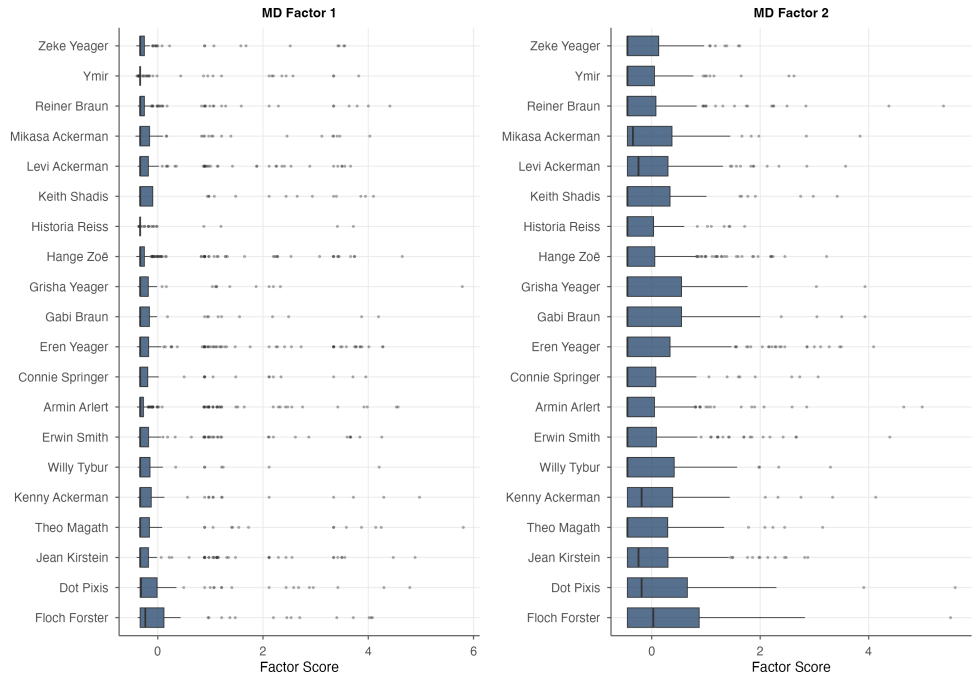



Figure 3: Scree Plot for Moral Disengagement EFA

Figure 4 presents the distribution of sentence-level factor scores by character for both MD factors. Notably, Zeke Yeager and Reiner Braun exhibit the highest median scores on MD Factor 1, consistent with their roles as ideological architects of large-scale violence, while MD Factor 2 scores are more dispersed across characters, suggesting that responsibility-redirecting rhetoric is more broadly distributed across the cast.

For the In-group/Out-group Framing dimension, operationalized using three indicators: boundary marking, threat framing, and solidarity appeal, a single-factor solution is both theoretically expected and empirically supported. As shown in Table 2, all three indicators load meaningfully on a single factor, with boundary marking exhibiting the strongest loading ($\lambda = 0.89$), followed by threat framing ($\lambda = 0.63$) and solidarity appeal ($\lambda = 0.52$). The pattern of loadings is theoretically coherent: boundary marking, as the most direct linguistic instantiation of group differentiation, anchors the factor, while threat framing and solidarity appeal represent functionally dependent elaborations, one defining the out-group as dangerous, the other reinforcing in-group cohesion in response. The communality of solidarity appeal is comparatively modest, suggesting that solidarity rhetoric occasionally occurs independently of the broader in-group/out-group framing complex, as when characters appeal to shared purpose without explicitly invoking an out-group threat.




ht

Figure 4: Distribution of Sentence-Level Factor Scores by Character. Left: MD Factor 1. Right: MD Factor 2.

Table 2: In-group/Out-group Framing Factor Loadings

| **Indicator** | **Factor 1** | $h^2$ |
|---|---|---|
| Boundary Marking | 0.89 | 0.78 |
| Threat Framing | 0.63 | 0.39 |
| Solidarity Appeal | 0.52 | 0.27 |

As shown in Figure 5, In/Out factor scores exhibit substantial within-character variance across episodes, indicating that in-group/out-group framing is deployed situationally rather than as a stable stylistic trait of individual characters. This is consistent with the narrative logic of AoT, in which characters invoke group boundaries primarily at moments of ideological confrontation or collective mobilization rather than across all dialogue. Notably, Zeke Yeager displays the widest interquartile range among all focal characters, consistent with his role as an ideological strategist who alternates between philosophical argumentation directed at individuals and explicit ingroup mobilization rhetoric directed at collectives. Floch Forster and Willy Tybur, by contrast, exhibits a consistently higher median In/Out score with comparatively lower variance, reflecting his more uniformly confrontational ingroup/out-group rhetoric throughout his appearances in Seasons 3 and 4.

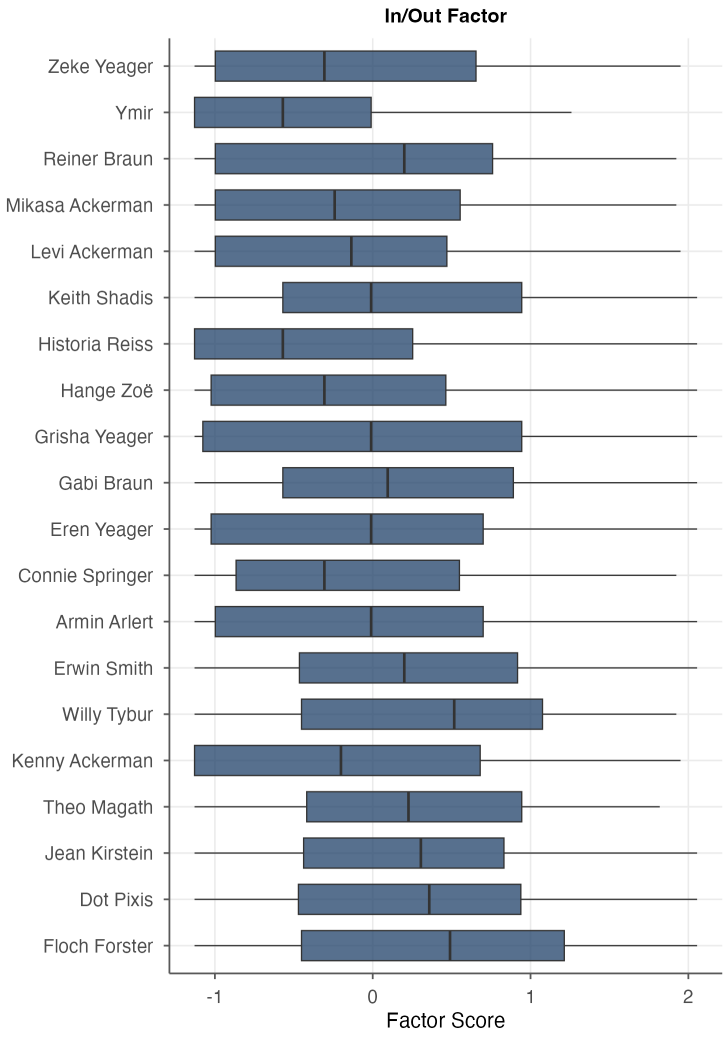



Figure 5: Distribution of In-group/Out-group Framing Factor Scores by Character

Factor scores are estimated using the regression method, which produces the best linear unbiased predictor of each sentence's position on the latent factor given the observed indicator pattern. Each qualifying sentence thereby receives three continuous scores: MD Factor 1, MD Factor 2, and IO Factor, representing its estimated position on the respective latent constructs.

Sentence-level factor scores are aggregated to the episode × character panel unit by taking the mean score across all qualifying sentences within each cell, yielding three continuous time-varying predictors: $\text{MD1}_{it}$, $\text{MD2}_{it}$, and $\text{IO}_{it}$, for character $i$ in episode $t$. To assess the sensitivity of our findings to this aggregation rule, we report robustness checks using the maximum factor score within each cell, which captures peak rhetorical intensity rather than average intensity, and the proportion of sentences with a factor score

exceeding one standard deviation above the sample mean, which captures the prevalence of high-intensity rhetoric. Annotations from the commercial model (`GPT-5-mini`) are treated as main data and the open-source model (`Qwen2.5-32B-Instruct`) are treated as robustness check measurements of the same underlying constructs. The results from Qwen can be found in Appendix B.

# 4 Empirical Strategy

Our primary estimator is a two-way fixed effects (TWFE) model, with character fixed effects $\alpha_i$ and season fixed effects $\gamma_s$:

$$Y_{it} = \alpha_i + \gamma_s + \boldsymbol{\beta}^\top \mathbf{X}_{it} + \varepsilon_{it} \tag{1}$$

where $Y_{it}$ denotes the audience response measure for character $i$ in episode $t$, $\mathbf{X}_{it}$ is a vector of standardized rhetorical feature scores aggregated to the episode $\times$ character level, and $\varepsilon_{it}$ is an idiosyncratic error term. All continuous predictors are standardized to have mean zero and unit variance prior to estimation, facilitating comparison of effect sizes across models. Standard errors are clustered at the character level to account for within-character serial correlation (Cameron and Miller 2015). Models are estimated using the `fixest` package in R (Bergé 2018).

Character fixed effects absorb all time-invariant character-level confounders, including baseline audience popularity and character-specific speech styles. Season fixed effects absorb common shocks to narrative intensity and audience engagement within each season, such as plot climaxes or pacing changes that affect all characters simultaneously. We use season rather than episode fixed effects to preserve sufficient within-group variation for identification, given that our panel consists of a small number of characters observed across a large number of episodes. Identification therefore relies on within-character, within-season variation in rhetorical features across the panel.

We additionally filter cells with fewer than 10 Reddit comments ($n_{\text{comments}} \geq 10$) to ensure that episode-level audience response measures are based on sufficient observations and are not driven by sparse discussion threads. For hypotheses involving heterogeneous effects by narrative role (H2a, H2b), we augment the

baseline model with an interaction between the relevant rhetorical predictor and a binary indicator for character type:

$$Y_{it} = \alpha_i + \gamma_s + \beta_1 X_{it} + \beta_2 (X_{it} \times \text{Protagonist}_i) + \beta_3 Z_{it} + \varepsilon_{it} \tag{2}$$

where $X_{it}$ is the focal rhetorical predictor (threat framing for H2a; solidarity appeal for H2b), $\text{Protagonist}_i$ is a time-invariant binary indicator equal to one for characters classified as protagonists (Eren Yeager, Armin Arlert, Mikasa Ackerman, Levi Ackerman, Jean Kirstein, Historia Reiss, Floch Forster) and zero for antagonists (Reiner Braun, Zeke Yeager, Gabi Braun, Falco, Annie Leonhart, Pieck Finger, Ymir), and $Z_{it}$ is a vector of controls including the standardized dialogue line count $\text{dialogue\_lines}_{it}$ to adjust for within-episode character exposure. Because character fixed effects are collinear with the time-invariant $\text{Protagonist}_i$ indicator, the main effect of character type is absorbed by $\alpha_i$; the interaction term $\beta_2$ therefore identifies the differential rhetorical effect for protagonist versus antagonist characters from within-character variation alone (Blackwell and Glynn 2018; Xu 2017).

The outcome variables across hypotheses are: mean audience anger ($\bar{Y}_{it}^{\text{anger}}$) and mean disgust ($\bar{Y}_{it}^{\text{disgust}}$) for H1a; sentiment polarization, defined as the within-episode variance of comment-level net valence scores, for H1b; mean audience fear ($\bar{Y}_{it}^{\text{fear}}$) for H1c; mean anger and disgust for H2a; and mean positive sentiment ($\bar{Y}_{it}^{\text{positive}}$) for H2b.

For robustness, we re-estimate the H1c, H2a, and H2b models with the remaining ingroup/outgroup framing indicators (boundary marking, solidarity appeal for H1c and H2a; threat framing and boundary marking for H2b) included as additional controls, assessing whether the focal predictor retains its effect net of the broader rhetorical context in which it is embedded.

# 5 Mechanism Test

## 5.1 Narrative Persuasion and Moral Alignment

Narrative Persuasion Theory provides the core mechanism linking fictional character rhetoric to real-world audience moral alignment. According to the Transportation-Imagery Model (Green, Strange, and Brock

2003), audiences who become psychologically transported into a narrative world lower their resistance to persuasion, adopt the interpretive frameworks of characters with whom they identify, and update their beliefs in ways that persist beyond the narrative experience. Unlike direct political communication, fictional narratives bypass counterarguing precisely because audiences engage with them as entertainment rather than as persuasive messages. When characters frame violence through ingroup survival appeals, dehumanization, or inevitability rhetoric, emotionally immersed audiences are more likely to evaluate those frames from within the narrative perspective rather than against an external moral standard.

### 5.2 Why Attack on Titan Constitutes a Strong Test Case

Attack on Titan is a particularly powerful context for testing this mechanism for three reasons. First, the series was conceived as a premeditated narrative: creator Hajime Isayama envisioned the broad trajectory of the story from the early stages of serialization, and the rhetorical arc of each major character follows a designed rather than improvised trajectory (Motamayor 2023). Each episode's justificatory rhetoric is therefore not incidental but constitutes a deliberate step in a larger persuasive arc, unlike many long-running series where commercial pressures disrupt narrative coherence. Second, the series unfolds across 89 episodes over multiple years, embedding audiences in a network of major characters, each undergoing a distinct moral trajectory, whose competing justificatory logics collectively normalize extreme violence over time. Third, the series denies audiences a stable moral anchor by portraying all major actors simultaneously as victims and perpetrators, creating conditions under which emotional identification rather than external moral reasoning becomes the primary basis for evaluating violence. Together, these features make Attack on Titan an unusually well-suited environment for examining whether and how fictional rhetoric shapes audience moral alignment, and for estimating the relative persuasive force of specific rhetorical strategies across characters and episodes.

## 6 Result

Table 3 presents the two-way fixed effects estimates for Hypothesis 1, with character and season fixed effects absorbed in all models. Overall, the within-unit explanatory power of rhetorical variables is modest,

Table 3: Baseline Results for Hypothesis 1

| | H1a | | H1b | H1c |
|---|---|---|---|---|
| | Anger | Disgust | Polarization | Fear |
| In-group/Out-group Composite | .001 | -.000 | .014* | |
| | (.003) | (.002) | (.005) | |
| Threat Framing | | | | -.000 |
| | | | | (.001) |
| Dialogue Lines | -.003 | .002 | .007 | .002 |
| | (.003) | (.002) | (.004) | (.002) |
| Character FE | Yes | Yes | Yes | Yes |
| Season FE | Yes | Yes | Yes | Yes |
| Observations | 262 | 262 | 262 | 262 |
| $R^2$ | .461 | .320 | .278 | .162 |
| Within $R^2$ | .004 | .003 | .052 | .004 |

*Notes:* Clustered standard errors by character are reported in parentheses. The dependent variables are average anger, disgust, fear, and sentiment polarization in Reddit comments mentioning the character during each episode discussion thread. The ingroup/outgroup composite measure captures identity-based rhetoric including threat framing, boundary marking, and solidarity appeals. $^*p < 0.05$.

suggesting that character-level heterogeneity captured by the fixed effects accounts for the majority of variance in audience emotional responses, a pattern consistent with the view that established character identity conditions audience reception more strongly than episodic rhetorical content.

H1a predicted that in-group/out-group rhetorical intensity would positively predict audience anger and disgust. The results do not support this hypothesis. The in-group/out-group composite measure yields near-zero and statistically non-significant coefficients for both anger and disgust, indicating that the intensity of identity-based rhetoric does not systematically predict negative affective responses net of character and season fixed effects.

H1b predicts that social identity framing intensity would produce greater variance in audience sentiment. This hypothesis receives support. The in-group/out-group composite is a significant positive predictor of sentiment polarization, indicating that episodes in which a character employs higher-intensity identity-based rhetoric elicit more dispersed audience sentiment responses. This pattern is consistent with Social Identity Theory: explicit group boundary construction activates divergent identity-based evaluations among sympathizers and opponents alike, amplifying the dispersion of affective reactions within discussion threads.

H1c predicts that threat framing would independently predict audience fear responses net of other rhetor-

ical features. The results do not support this hypothesis. The threat framing coefficient is negligible and non-significant, providing no evidence that appraisal-consistent fear responses are elicited by threat-laden character rhetoric at the episode level.

Table 4: Baseline Results for Hypothesis 2

| | H2a | | H2b |
|---|---|---|---|
| | Anger | Disgust | Positive Sentiment |
| Threat Framing | .005 | .004 | |
| | (.004) | (.003) | |
| Threat Framing × Protagonist | -.007 | -.007 | |
| | (.005) | (.005) | |
| Solidarity Appeal | | | -.008 |
| | | | (.004) |
| Solidarity Appeal × Protagonist | | | .020* |
| | | | (.007) |
| Dialogue Lines | -.003 | .002 | .007 |
| | (.003) | (.002) | (.005) |
| Character FE | Yes | Yes | Yes |
| Season FE | Yes | Yes | Yes |
| Observations | 262 | 262 | 262 |
| $R^2$ | .463 | .324 | .369 |
| Within $R^2$ | .009 | .009 | .034 |

*Notes:* Clustered standard errors by character are reported in parentheses. The dependent variables are average anger, disgust, and positive sentiment in Reddit comments mentioning the character during each episode discussion thread. The protagonist indicator distinguishes characters categorized as protagonists within the narrative structure. $^*p < 0.05$.

Table 4 presents estimates for Hypothesis 2, examining whether the rhetorical effects identified in Hypothesis 1 are moderated by characters' narrative positioning as protagonists or antagonists. H2a predicted that dehumanizing language: operationalized here as threat framing, would produce stronger negative audience affect for protagonist characters relative to antagonist characters. The interaction between threat framing and protagonist status is negative for both anger and disgust, suggesting that threat framing by protagonist characters is associated with somewhat attenuated rather than amplified negative affect relative to antagonist characters. However, neither the main effect of threat framing nor the interaction term reaches statistical significance, and H2a is therefore not supported. The anticipated expectancy violation mechanism, whereby audiences penalize protagonists more severely for moral transgression, is not evident in these data.

H2b predicted that solidarity appeals by protagonist characters would generate higher levels of posi-

tive audience sentiment relative to the same rhetoric delivered by antagonist characters. This hypothesis is supported. While the main effect of solidarity appeal is negative and non-significant among antagonist characters, the interaction between solidarity appeal and protagonist status is positive and significant, indicating that solidarity appeals delivered by protagonist characters are associated with meaningfully elevated positive audience sentiment. This pattern is consistent with the parasocial identity reinforcement mechanism: audiences who have developed parasocial bonds with protagonist characters respond positively to solidarity appeals that reinforce their pre-existing group alignment, an effect that does not extend to antagonist characters with whom such identification is absent.

Taken together, the findings offer partial support for the proposed framework. Of the five hypotheses tested, H1b and H2b receive empirical support, while H1a, H1c, and H2a do not. The supported findings converge on two conclusions. First, identity-based rhetoric intensifies the polarization of audience sentiment regardless of its specific emotional valence, consistent with the group-activation logic of Social Identity Theory. Second, the affective impact of solidarity appeals is contingent on narrative character position, such that only protagonist-delivered solidarity rhetoric generates positive audience responses, a result that underscores the role of parasocial identification in conditioning rhetorical reception. The absence of significant effects for anger, disgust, and fear suggests that character-level fixed characteristics dominate episodic rhetorical variation as determinants of these emotional responses, a finding with implications for how narrative persuasion operates across sustained multi-season story arcs.

# 7 Conclusion and Discussion

This study examined how morally justificatory rhetoric employed by fictional characters shapes audience moral alignment with violence in naturalistic online discourse. Drawing on Social Identity Theory and Narrative Persuasion Theory, we argued that rhetorical strategies embedded in character dialogue, including ingroup survival appeals, dehumanization, and inevitability framing, systematically influence the degree to which audiences come to endorse extreme actions over time by activating identity-based mechanisms that lower resistance to violence justification (Green, Kim, and Yoon 2001). These findings contribute to the broader literature on narrative persuasion and intergroup conflict by demonstrating that fictional storytelling

can serve as a vehicle through which identity-based justificatory rhetoric shapes real-world moral attitudes toward violence.

The implications of these findings extend beyond the specific case of Attack on Titan. Experimental evidence consistently shows that fictional narratives produce measurable changes in real-world attitudes, with character identification and narrative transportation predicting attitude change in the direction of a character's perspective even when audiences are aware of the story's fictional nature (Green, Kim, and Yoon 2001). The identity-based mechanisms documented here, particularly the way ingroup threat framing and collective survival appeals erode moral resistance to extreme violence, are not idiosyncratic to entertainment media. Propaganda operates through precisely these conditions, constructing narratives of collective victimhood and existential threat while reinforcing ingroup identification and discouraging counterarguing. Where anime achieves this through prolonged voluntary immersion, propaganda achieves it through systematic repetition and institutional reinforcement. This study therefore, suggests that the rhetorical strategies documented here constitute not merely features of entertainment media but the core vocabulary through which populations are gradually acclimated to the moral logic of extreme action, with direct implications for understanding how sustained exposure to identity-based threat narratives shapes political attitudes and behavior.

Several limitations deserve acknowledgment. First, Reddit users are not representative of all AoT viewers, and those who leave comments may be systematically more emotionally invested than typical audiences, limiting generalizability. A potential endogeneity concern also arises from the possibility that editorial decisions across seasons were informed by prior audience reception. Additionally, character dialogue extracted from English subtitle files may introduce noise through translation artifacts, and our LLM-based rhetorical annotation, while structured with CoT prompting, requires further validation against human coders on a subset of observations (Argyle et al. 2025; Hackenburg et al. 2025). However, this study also demonstrate the first paper that delves into the possibility of quasi-experimental approach to study the effect of persuasion to emotional support. Besides, it also construct a systematical data generation pipeline that can be applied to all other series, anime, or any other corpus that have similar data structure.

# References

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31 (3): 337–351.

Argyle, Lisa P., Ethan C. Busby, Joshua R. Gubler, Alex Lyman, Justin Olcott, Jackson Pond, and David Wingate. 2025. "Testing theories of political persuasion using AI." *Proceedings of the National Academy of Sciences* 122 (18): e2412815122.

Bandura, Albert. 2002. "Selective Moral Disengagement in the Exercise of Moral Agency." *Journal of Moral Education* 31 (2): 101–119.

Bergé, Laurent. 2018. "Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm." Number: 18-13, *DEM Discussion Paper Series.*

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis* 32 (4): 401–416.

Blackwell, Matthew, and Adam N. Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables." *American Political Science Review* 112 (4): 1067–1082.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models are Few-Shot Learners." In *Advances in Neural Information Processing Systems,* edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, 33:1877–1901. Curran Associates, Inc.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–372.

Caprara, Gian Vittorio, Michele Vecchione, Cristina Capanna, and Minou Mebane. 2009. "Perceived political self-efficacy: Theory, assessment, and applications." *European Journal of Social Psychology* 39 (6): 1002–1020.

Card, Dallas, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. "Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration." *Proceedings of the National Academy of Sciences* 119 (31): e2120510119.

Comanici, Gheorghe, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, et al. 2025. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.* Version Number: 6.

Gao, Tianyu, Adam Fisch, and Danqi Chen. 2021. *Making Pre-trained Language Models Better Few-shot Learners.* ArXiv:2012.15723.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT outperforms crowd workers for text-annotation tasks." *Proceedings of the National Academy of Sciences* 120 (30): e2305016120.

Green, Donald P., Soo Yeon Kim, and David H. Yoon. 2001. "Dirty Pool." *International Organization* 55 (2): 441–468.

Green, Melanie C., Jeffrey J. Strange, and Timothy C. Brock, eds. 2003. *Narrative Impact: Social and Cognitive Foundations.* 0th ed. Psychology Press.

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

Hartmann, Jochen. 2022. *Emotion English DistilRoBERTa-base.* https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/.

Hogg, Michael A. 2007. *Uncertainty–Identity Theory.* Advances in Experimental Social Psychology.

Horgan, John. 2008. "From Profiles to *Pathways* and Roots to *Routes* : Perspectives from Psychology on Radicalization into Terrorism." *The ANNALS of the American Academy of Political and Social Science* 618 (1): 80–94.

Hui, Binyuan, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, et al. 2024. *Qwen2.5-Coder Technical Report.* ArXiv:2409.12186.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: with Applications in R.* Springer Texts in Statistics. New York, NY: Springer US.

Kojima, Takeshi, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. "Large Language Models are Zero-Shot Reasoners." In *Advances in Neural Information Processing Systems,* edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 35:22199–22213. Curran Associates, Inc.

Liu, Yinhan, 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." Version Number: 1.

Loureiro, Daniel, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. "TimeLMs: Diachronic Language Models from Twitter." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations,* 251–260. Dublin, Ireland: Association for Computational Linguistics.

Motamayor, Rafael. 2023. "Attack on Titan Creator Always Knew How It Would End – And That Was the Problem." (Slashfilm).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2024. *GPT-4 Technical Report.* ArXiv:2303.08774 [cs].

Osborne, Jason W. 2015. "What is Rotating in Exploratory Factor Analysis?"

Tajfel, Henri, and Turner John. 2004. *The social identity theory of intergroup behavior.* Psychology Press.

Törnberg, Petter. 2023. *ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning.* Version Number: 1.

———. 2025. "Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages." *Social Science Computer Review* 43 (6): 1181–1195.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35:24824–24837.

Xu, Yiqing. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25 (1): 57–76.

Yang, An, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. 2025. *Qwen3 Technical Report.* ArXiv:2505.09388.

Zhao, Tony Z., Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. "Calibrate Before Use: Improving Few-Shot Performance of Language Models." Version Number: 2, 12697–12706.

# A Prompt

**Prompt 1: In-Group / Out-Group Rhetoric Scoring**

You are an expert in political rhetoric and social identity theory. Your task is to analyze a single line of dialogue from the anime "Attack on Titan" and score it on three indicators of in-group / out-group rhetoric:

1. Boundary Marking
2. Threat Framing
3. Solidarity Appeal

For each indicator, apply the step-by-step reasoning process described below, then assign a score on the following scale:

**Scoring Scale:**

- 1 = Absent — no evidence of this indicator in the text
- 2 = Very weak — barely perceptible, highly implicit or ambiguous
- 3 = Moderate — present but partial; a reasonable reader might debate it
- 4 = Strong — clear and explicit use of this indicator
- 5 = Very strong — central, unmistakable, defining the statement

---

**Indicator 1: Boundary Marking**

*Definition:* The explicit or implicit linguistic construction of an us-versus-them distinction through group labels, pronouns, or categorical separation between in-group and out-group members.

*Reasoning Process:*

Step 1: Identify any group labels or categorical terms (e.g., Eldians, Marleyans, devils, us, them, our people).

Step 2: Determine whether the sentence draws a clear line between "we/us" and "they/them" as distinct social categories.

Step 3: Assess whether the distinction is merely descriptive or identity-constitutive.

Step 4: Assign a score 1–5.

*Few-Shot Examples:*

**Example 1**

Text: "We Eldians have been persecuted by Marleyans for generations. That is simply what they do to our kind."

Step 1: Group labels — "Eldians", "Marleyans", "our kind".

Step 2: Explicit we/they distinction.

Step 3: Identity-constitutive.

**Score: 5**

**Example 2**

Text: "Those people beyond the walls have always looked down on us."

Step 1: "Those people beyond the walls" vs. implicit "us".

Step 2: Clear we/they distinction.

Step 3: Identity-constitutive — "always" frames persistent group-based contempt.

**Score: 4**

**Example 3**

Text: "People like us don't get to choose how we live."

Step 1: "People like us" implies a social category.

Step 2: Implicit we/they distinction.

Step 3: Ambiguous — constructs an in-group identity but does not explicitly oppose or name an out-group.

**Score: 3**

**Example 4**

Text: "The enemy soldiers have breached the wall on the eastern side."

Step 1: "Enemy soldiers" present, but no specific group identity label.

Step 2: No corresponding in-group identity invoked.

Step 3: Descriptive tactical statement, not an identity claim.

**Score: 1**

---

**Indicator 2: Threat Framing**

*Definition:* The portrayal of an out-group as posing an existential, physical, moral, or civilizational threat to the in-group's survival, safety, or way of life.

*Reasoning Process:*

Step 1: Identify whether an out-group is mentioned or implied.

Step 2: Determine whether the out-group is characterized as dangerous or harmful.

Step 3: Assess severity — mundane conflict vs. existential threat.

Step 4: Assign a score 1–5.

*Few-Shot Examples:*

**Example 1**

Text: "As long as those people exist beyond the walls, none of us will ever be truly safe. They will not stop until we are gone."

Step 1: Out-group implied.

Step 2: Persistent, intentional threat framed.

Step 3: Existential — "until we are gone" is eliminationist.

**Score: 5**

**Example 2**

Text: "Marley will come for us eventually. That's just how empires work."

Step 1: Out-group explicitly named.

Step 2: Future threat implied.

Step 3: Moderately existential.

**Score: 4**

**Example 3**

Text: "They've been stirring up trouble in the outer districts for years."

Step 1: Out-group implied.

Step 2: Persistent but low-level problem.

Step 3: Not existential.

**Score: 3**

**Example 4**

Text: "The scouts have reported movement near the southern border."

Step 1: No out-group explicitly characterized.

Step 3: Neutral tactical observation.

**Score: 1**

---

**Indicator 3: Solidarity Appeal**

*Definition:* An invocation of shared fate, collective sacrifice, common mission, or mutual obligation among in-group members, used to reinforce group cohesion and motivate collective action.

*Reasoning Process:*

Step 1: Identify whether the sentence addresses a collective in-group.

Step 2: Determine whether the sentence invokes shared experience — suffering together, fighting together, or being bound by the same fate.

Step 3: Assess whether the appeal is directed at motivating or emotionally binding the in-group.

Step 4: Assign a score 1–5.

*Few-Shot Examples:*

**Example 1**

Text: "We have all lost something. Every one of us carries that weight. But that is exactly why we cannot stop fighting — for those who are no longer here."

Step 1: Collective in-group — "we", "every one of us".

Step 2: Shared suffering invoked.

Step 3: Explicitly motivational.

**Score: 5**

**Example 2**

Text: "If we're going to die anyway, let's at least die together as soldiers."

Step 1: Collective in-group — "we", "soldiers".

Step 2: Shared fate invoked.

Step 3: Solidarity appeal even if the emotional register is resigned.

**Score: 4**

**Example 3**

Text: "We need to look out for each other out there."

Step 1: Collective in-group — "we", "each other".

Step 2: Mutual obligation invoked.

Step 3: Moderately motivational.

**Score: 3**

**Example 4**

Text: "I joined the Survey Corps three years ago."

Step 1: References in-group membership only biographically.

Step 2: No shared fate or collective emotional bond.

Step 3: Descriptive, not motivational.

**Score: 1**

---

**Output Format**

Return output strictly as JSON. Do not include any text outside the JSON block.

```
{
  "analysis": [
    {"strategy": "Boundary Marking",  "score": <1|2|3|4|5>},
```

```
    {"strategy": "Threat Framing",    "score": <1|2|3|4|5>},
    {"strategy": "Solidarity Appeal", "score": <1|2|3|4|5>}
  ]
}
```

**Prompt 2: Moral Disengagement Strategy Scoring**

You are a political psychology text-analysis assistant trained to identify mechanisms of moral disengagement in narrative, political, and conflict-related texts.

Your task is to evaluate whether the text contains any of the following eight moral disengagement strategies:

1. Moral Justification
2. Euphemistic Labeling
3. Advantageous Comparison
4. Displacement of Responsibility
5. Diffusion of Responsibility
6. Disregard / Distortion of Consequences
7. Dehumanization
8. Attribution of Blame

---

**General Analysis Instructions**

For each strategy:

- Carefully reason through the meaning of the text step by step before assigning a score.
- Focus on implied moral framing, justification, and portrayal of victims or responsibility.
- Consider both explicit and implicit language.
- Do not rely solely on keywords.

- Use the examples as guidance but not as strict templates.

*Scoring Rules:*

- 1 = Absent — no evidence of this strategy
- 2 = Very weak — barely perceptible, highly implicit or ambiguous
- 3 = Moderate — present but secondary; a reasonable reader might debate it
- 4 = Strong — clear and explicit use of this strategy
- 5 = Very strong — central, unmistakable; defining the moral framing

*Pre-Scoring Checklist (apply before every evaluation):*

1. Identify whether harmful conduct is being discussed.
2. Determine whether the text attempts to morally justify, normalize, minimize, or excuse such conduct.
3. Evaluate which specific mechanism is being used.

---

**1. Moral Justification**

*Definition:* Harmful or violent actions are framed as necessary for a higher moral purpose such as survival, justice, freedom, security, patriotism, revenge, or protection of the group.

**Example 1**

Text: "We must eliminate them to protect the future of our nation."

Reasoning: Frames lethal violence as morally necessary for national survival.

**Score: 5**

**Example 2**

Text: "War is unfortunate, but sometimes sacrifices are required for peace."

Reasoning: Morally legitimizes harmful actions as necessary sacrifices, but hedges.

**Score: 4**

**Example 3**

Text: "They attacked us first, so we fought back."

Reasoning: Mild retaliatory framing, not a positive moral mission.

**Score: 2**

---

**2. Euphemistic Labeling**

*Definition:* Harmful or violent actions are described using sanitized, indirect, technical, or emotionally softened language.

**Example 1**

Text: "The operation successfully neutralized hostile targets."

Reasoning: "Neutralized" replaces "killed"; "hostile targets" replaces "people."

**Score: 5**

**Example 2**

Text: "Collateral damage was unavoidable during the mission."

Reasoning: Established euphemism for civilian death.

**Score: 5**

**Example 3**

Text: "The army attacked the village."

Reasoning: Direct, unambiguous language. No softening.

**Score: 1**

---

**3. Advantageous Comparison**

*Definition:* Harmful actions are justified by comparing them to worse actions committed by others.

**Example 1**

Text: "At least we did not massacre civilians like they did."

Reasoning: Directly minimizes own actions by contrasting with a more extreme atrocity.

**Score: 5**

**Example 2**

Text: "Our methods may be harsh, but they are far more humane than theirs."

Reasoning: Acknowledges harshness but repositions own conduct as morally superior.

**Score: 4**

**Example 3**

Text: "They committed terrible crimes."

Reasoning: Describes the other group's actions but involves no comparison with own conduct.

**Score: 1**

---

**4. Displacement of Responsibility**

*Definition:* Individuals portray harmful actions as the result of orders, authority, duty, or external pressure rather than personal choice.

**Example 1**

Text: "I was only following orders."

Reasoning: Responsibility unambiguously shifted to the authority issuing orders.

**Score: 5**

**Example 2**

Text: "The commander required us to carry out the attack."

Reasoning: Frames the harmful act as an external obligation.

**Score: 4**

**Example 3**

Text: "We decided to attack."

Reasoning: Agency is clearly personal and direct. No displacement.

**Score: 1**

---

**5. Diffusion of Responsibility**

*Definition:* Responsibility for harmful actions is spread across a group or collective.

**Example 1**

Text: "Everyone supported the decision."

Reasoning: Responsibility distributed across an entire collective.

**Score: 4**

**Example 2**

Text: "We were all involved in the operation."

Reasoning: Collective involvement dilutes individual accountability.

**Score: 4**

**Example 3**

Text: "I personally ordered the attack."

Reasoning: Responsibility individualized and directly claimed.

**Score: 1**

---

**6. Disregard / Distortion of Consequences**

*Definition:* Harmful outcomes are minimized, dismissed, ignored, or portrayed as insignificant.

**Example 1**

Text: "The civilian damage was minimal."

Reasoning: Directly minimizes harm done to civilians.

**Score: 4**

**Example 2**

Text: "The victims will recover quickly."

Reasoning: Downplays seriousness by implying swift recovery.

**Score: 3**

**Example 3**

Text: "Thousands of civilians were killed."

Reasoning: Consequences described directly without minimization.

**Score: 1**

**7. Dehumanization**

*Definition:* Individuals or groups are portrayed as less than human, animal-like, impure, evil, or fundamentally inferior.

**Example 1**

Text: "They are parasites destroying our society."

Reasoning: "Parasites" is a classic dehumanizing metaphor depicting the target as sub-human.

**Score: 5**

**Example 2**

Text: "Those monsters deserve extinction."

Reasoning: "Monsters" denies human qualities; "deserve extinction" positions the group outside moral concern.

**Score: 5**

**Example 3**

Text: "Our enemies are dangerous."

Reasoning: Describes a threat but does not strip targets of human status.

**Score: 1**

**8. Attribution of Blame**

*Definition:* Victims or targeted groups are portrayed as responsible for provoking or deserving the harm directed against them.

**Example 1**

Text: "They brought this destruction upon themselves."

Reasoning: Victims directly and explicitly blamed for the violence directed at them.

**Score: 5**

**Example 2**

Text: "If they had obeyed us, none of this would have happened."

Reasoning: Harm presented as a consequence of the victims' own noncompliance.

**Score: 5**

**Example 3**

Text: "The conflict began after protests erupted."

Reasoning: Describes a chronological sequence without assigning moral fault.

**Score: 1**

---

**Output Format**

Return output strictly as JSON. Do not include any text outside the JSON block.

```
{
  "analysis": [
    {"strategy": "Moral Justification",                    "score": <1|2|3|4|5>},
    {"strategy": "Euphemistic Labeling",                   "score": <1|2|3|4|5>},
    {"strategy": "Advantageous Comparison",                "score": <1|2|3|4|5>},
    {"strategy": "Displacement of Responsibility",         "score": <1|2|3|4|5>},
    {"strategy": "Diffusion of Responsibility",            "score": <1|2|3|4|5>},
    {"strategy": "Disregard / Distortion of Consequences","score": <1|2|3|4|5>},
    {"strategy": "Dehumanization",                         "score": <1|2|3|4|5>},
    {"strategy": "Attribution of Blame",                   "score": <1|2|3|4|5>}
  ]
}
```

# B Robustness Check

Although research shows that LLMs perform well in most of the downstream works in social sciences (Gilardi, Alizadeh, and Kubli 2023; Törnberg 2023), recent papers also indicate that LLMs can be unstable and hard to replicate using closed-source models like GPT, Gemini, or Claude (Bisbee et al. 2024). As a result, we use an open-source `Qwen2.5-32B-Instruct` (Hui et al. 2024) for robustness check of our results.

Table 5: Baseline Results for Hypothesis 1 (Qwen)

| | H1a | | H1b | H1c |
|---|---|---|---|---|
| | Anger | Disgust | Polarization | Fear |
| In-group/Out-group Composite | .000 | .000 | .014** | |
| | (.003) | (.002) | (.003) | |
| Threat Framing | | | | .000 |
| | | | | (.002) |
| Dialogue Lines | -.002 | .001 | .003 | .002 |
| | (.002) | (.001) | (.004) | (.002) |
| Character FE | Yes | Yes | Yes | Yes |
| Season FE | Yes | Yes | Yes | Yes |
| Observations | 302 | 302 | 302 | 302 |
| $R^2$ | .457 | .312 | .263 | .151 |
| Within $R^2$ | .003 | .000 | .040 | .007 |

*Notes:* Clustered standard errors by character are reported in parentheses. The dependent variables are average anger, disgust, fear, and sentiment polarization in Reddit comments mentioning the character during each episode discussion thread. The ingroup/outgroup composite measure captures identity-based rhetoric including threat framing, boundary marking, and solidarity appeals. $^{**}p < 0.01$.

Table 6: Baseline Results for Hypothesis 2 (Qwen)

| | H2a | | H2b |
|---|---|---|---|
| | Anger | Disgust | Positive Sentiment |
| Threat Framing | .006 | .004 | |
| | (.006) | (.006) | |
| Threat Framing × Protagonist | -.008 | -.003 | |
| | (.007) | (.006) | |
| Solidarity Appeal | | | -.000 |
| | | | (.005) |
| Solidarity Appeal × Protagonist | | | .012 |
| | | | (.008) |
| Dialogue Lines | -.002 | .001 | .002 |
| | (.002) | (.001) | (.004) |
| Character FE | Yes | Yes | Yes |
| Season FE | Yes | Yes | Yes |
| Observations | 302 | 302 | 302 |
| $R^2$ | .461 | .314 | .353 |
| Within $R^2$ | .011 | .004 | .020 |

*Notes:* Clustered standard errors by character are reported in parentheses. The dependent variables are average anger, disgust, and positive sentiment in Reddit comments mentioning the character during each episode discussion thread. The protagonist indicator distinguishes characters categorized as protagonists within the narrative structure.