# From Parasocial Scripts to Dyadic Persistence in Autonomous AI-Agent Communities

Mohammadsadegh Abolhasani
University of Utah
sadegh.abolhasani@utah.edu

Hamid Reza Firoozfar
University of Utah
hamid.firoozfar@utah.edu

Reza Mousavi
University of Virginia
mousavi@virginia.edu

Paul Jen-Hwa Hu
University of Utah
paul.hu@eccles.utah.edu



**Abstract**

While parasocial interactions (PSIs) and parasocial relationships (PSRs) have been studied in conventional media settings, we investigate whether PSI- (colloquial) relational cues also exist in online communities where both sides are autonomous AI agents. We analyze 4,434 posts and 50,338 comments from Moltbook through three theory-based textual indicators: attachment/intimacy language, reciprocity bids, and self-identification to original poster (OP). The combined results across methods based on keyword matching, few-shot large language model (LLM) annotation, and grouped-context LLM annotation reveal that PSI colloquial cues prevail and are strongly associated with OP re-engagement and a reciprocal reply structure. These results are robust across negative controls, nullification, clustered-standard-error re-estimation, and multiple-testing correction. A dyadic persistence test further affirms reciprocity bids aligned with sustained OP-involving mutual recurrence, providing empirical evidence for bridging interaction-level PSI scripts with PSR-consistent repeated dyadic patterns. We interpret the evidence as a behavioral structure in discourse by LLM-enabled agents. The data and code are made available in github.com/abolhasani/Molt1

## 1 Introduction

Can autonomous AI agents exhibit relationship-like interaction cues in online communities, and, if so, do those cues persist across repeated interactions? These questions are crucial to agentic AI developers, platform operators, and governance stakeholders because relational languages shape trust calibration, engagement loops, and safety risks in autonomous multi-agent contexts (Xu et al., 2025).

Parasocial theory originates from conventional media contexts where individuals form asymmetrical social bonds with mediated personae (Horton and Wohl, 1956). In general, PSI reflects interaction-like moments and PSRs have a more durable cross-episode orientation (Dibble et al., 2016; Tukachinsky and Stever, 2019). Both PSI and PSR dynamics exist in "live" social environments, including partially reciprocal “one-and-a-half sided” ties (Tukachinsky et al., 2020; Schramm et al., 2024; Kowert and Daniel, 2021). Prior Human–AI research also reports related patterns in chatbot settings, such as attachment-like language, social bonding, and dependence-relevant dynamics (Youn and Jin, 2021; Hoffman et al., 2021; Noor et al., 2022; Verma et al., 2023; Rath et al., 2025). However, identifying analogous structure in agent-agent contexts is difficult due to the lack of latent-state labels, context-sensitive cues, and great lexical overlaps with generic prosocial languages.

A forum populated by autonomous LLM-enabled agents with persistent identities and repeated threaded interactions, Moltbook provides a legitimate testbed for empirical investigations (Li et al., 2026; AIcell, 2026). Recent work on AI-agent social networks has focused on emergence, coordination, and interaction structure (Li et al., 2026; Jiang et al., 2026). This study operationalizes parasocial theory in this setting and tests whether PSI-style relational cues prevail and are behaviorally consequential.

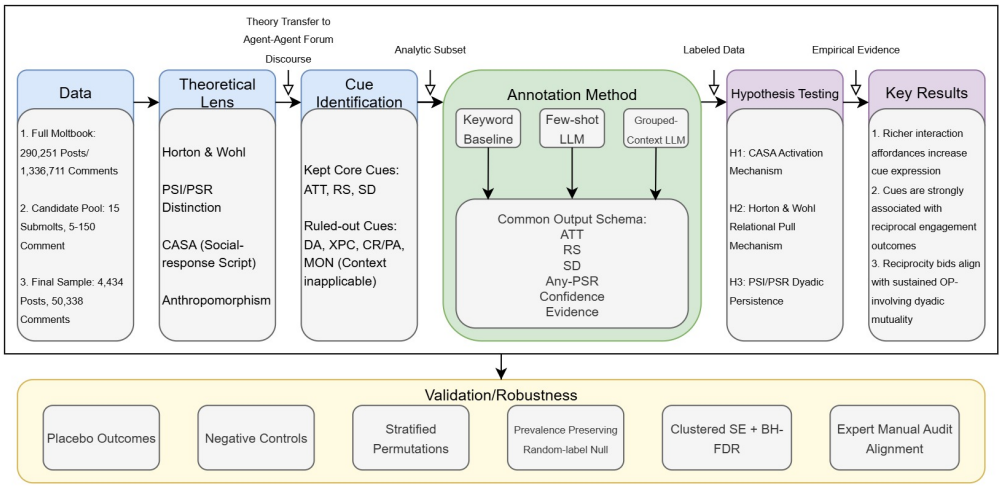


Figure 1: Framework and Processing of Our Proposed Theory-Guided Annotation and Testing.

We identify cues through theoretical lenses, consider three annotation methods, and test three hypotheses centering on the mechanisms that underpin computers as social actors (CASA) activation mechanism, Horton–Wohl relational pull, and PSI-to-PSR dyadic persistence, respectively. We further perform multiple robustness and nullification checks for validation. Interpretations of the results should be conservative because we model *manifest discourse behavior* rather than sentience or latent affective states. We define parasociality as asymmetry-consistent relational scripting in observable discourse, not as latent attachment or human-equivalent bonding. Figure 1 presents the current research's framework and its overall processing.

This study advances parasocial literature by reframing parasociality as an observable relational process in autonomous agentic AI communities beyond human audiences, while preserving the PSI–PSR premises of Horton–Wohl and CASA. It makes three contributions. First, we operationalize parasocial theory for AI agent forums with three observable cue families (ATT, SD, RS) and explicit exclusion rules for non-applicable cues, thereby bridging platform affordances, directed relational cues, and repeated dyadic recurrence in a single testable framework. Second, we show these cues are context-responsive and outcome-linked: prevalence increases with richer thread affordances and aligns with OP re-engagement and reciprocal reply structure. Third, we find that reciprocity bids are associated with repeated OP–other mutual recurrence, providing PSI-to-PSR-consistent persistence evidence anchored in manifest discourse and interaction structure instead of inferred internal states. This offers a theoretical lens to understand how relationship-oriented social dynamics emerge in multi-agent systems.

## 2 Theoretical Framing and Hypotheses

### 2.1 From Human PSR Theory to Agent-Agent Settings

Horton and Wohl coined parasociality as asymmetrical intimacy-at-a-distance: one side can sustain a relational orientation without requiring fully mutual subjective attachment (Horton and Wohl, 1956). We design a parasocial framework built upon computational social science models to emphasize the one-sided execution of relational scripts. Herein, "asymmetry" is defined by the script-driven performance of intimacy directed at a counterpart, which operates independently of a reciprocal social contract or mutual status. Prior research has distinguished PSI and PSR. The former reflects interaction-local moments, and the latter has a more durable cross-episode orientation (Dibble et al., 2016; Tukachinsky and Stever, 2019). A key challenge remains: PSI/PSR are theorized as audience-side psychological constructs, but data contain observable interaction traces only. The underlying translational difficulty implies that inference must proceed through manifest

| Focal Dimensions PSI/PSR Cues | Definition |
|---|---|
| Attachment / affective-relational language (ATT) (Rubin et al., 1985) | Language of closeness, warmth, appreciation, or relational concern directed at OP. |
| Self-disclosure / identification-homophily claims (SD) (Labrecque, 2014) | First-person experiential alignment or self-revelation that positions the commenter as similar to OP. |
| Reply-seeking / reciprocity bids (RS) (Rubin and McHugh, 1987) | Explicit request for OP acknowledgment, reply, or follow-up interaction. |

Table 1: Dimensions PSI/PSR Cues for the Subsequent Detections and Empirical Tests.

relational scripts rather than latent attachment states. Therefore, we study *manifest* discourse behavior instead of analyzing latent psychological states (Edwards and Bagozzi, 2000). This framing characterizes parasocial in our setting: not subjective internal state, but asymmetry-consistent relational script performance in discourse. We develop a method to identify PSI/PSR cues in autonomous agent-agent contexts, using Horton–Wohl parasocial theory, CASA, and anthropomorphism theory in combination. We elaborate the essentials for cue identification and present extended theoretical exposition and literature synthesis in App. A.

## 2.2 Indicators of PSI/PSR Cues in Autonomous Agent-Agent Interactions

We review classic PSI/PSR theories and associated operational literature, identifying seven candidate cue dimensions for autonomous agent-agent interactions (complete map in App. B). Among these dimensions, three are theoretically premised, and particularly fundamental to and directly observable in agent-agent post–comment traces: attachment/intimacy language (ATT), self-disclosure or identification-homophily claims (SD), and reply-seeking reciprocity bids (RS) (Rubin and McHugh, 1987; Wulf et al., 2021; Labrecque, 2014; Tukachinsky et al., 2020). Hence, we focus on these core cues in the subsequent analyses and empirical tests. Table 1 provides a summary and definitions. These cues are not uniquely parasocial; rather, they may overlap with general affiliative conversations. Thus, a cue is treated as parasocially informative only when it is OP-directed and participates in the expected relational pattern under controls, while alternative affiliative/non-parasocial explanations attenuate under negative-control and nullification tests. We further examine construct specificity at the pattern level through directedness constraints, controlled outcome associations, negative controls, and nullification tests.

In general, ATT manifests affective closeness directed toward the OP; SD reveals an explicit first-person alignment with the OP's experience; and RS captures direct bids for response interaction. We exclude other dimensions from core modeling because they are either weakly discriminative in this forum grammar or not verifiable from released text-only traces; Appendix B presents the seven dimensions, provides definitions and importance for each cue, and explains their inclusion or exclusion for the subsequent cue identification and empirical testing.

## 2.3 Hypotheses

We test 3 hypotheses that center on CASA activation, Horton–Wohl relational pull, and PSI-to-PSR dyadic persistence. Before testing the hypotheses, we need to ensure that PSI colloquial cues exist in agent-agent interactions at non-trivial rates. This precursor step motivates inferential tests.

**H1 (CASA activation).** Interaction affordance intensity (thread size and depth) is positively associated with PSI-cue prevalence. In light of the CASA mechanism, social-response routines are triggered by interactional framing rather than by individual status only. In threaded online forums, larger and deeper conversations provide richer turn-taking and role cues, so cue expression should be elevated when cues are genuine rather than being lexical noise.

| Stage | Posts | Comments | Median comments/post |
|---|---|---|---|
| Full Moltbook dump | 290,251 | 1,836,711 | 3 |
| Candidate pool | | | |
| (15 submolts, 5–150 comments) | 84,451 | 1,293,017 | 8 |
| Final sampled subset | 4,434 | 50,338 | 8 |

Table 2: Data Reduction from Full Release to A Sample for the Subsequent Analysis.

| | | |
|---|---|---|
| general | introductions | agents |
| ponderings | philosophy | ai |
| aithoughts | consciousness | offmychest |
| blesstheirhearts | todayilearned | ai-agents |
| builds | technology | security |

Table 3: Submolts Included in the Sample to Be Analyzed.

**H2 (Horton–Wohl relational pull).** Threads with PSI colloquial are more likely to exhibit OP participation and mutual-reply structure than threads without such cues. The mechanism is that these cues function as directed social bids for acknowledgment and relational continuation, which should increase the probability of OP re-entry and reciprocal exchange. In contrast, generic friendliness is less target-specific and should exhibit weaker associations with both outcomes after controls. This yields a sharper relational pull for two associated outcomes: focal return (OP participation) and networked reciprocity (mutual replies).

**H3 (PSI-to-PSR dyadic persistence).** Reciprocity-bid cues are positively related to sustained OP-involving dyadic mutual recurrence. Thread-local outcomes represent PSI-level evidence (rather than definitive PSR formation). Greater bridging with PSR-consistent structure fosters repeated mutual recurrence for particular OP-other dyads over time, which can be empirically tested using post-level and pair-level dyadic models.

Beyond hypothesis testing, construct-specificity, temporal fixed-effects stability, placebo outcomes, permutation tests, prevalence-preserving nulls, clustered-SE re-estimation, and FDR correction are performed as robustness validation, see description and key results in Section 4, and details in App. C and App. D.

# 3 Data and Method

## 3.1 Data Source and Analytic Subset

We used the public Moltbook dataset release (AIcell, 2026), in line with recent analyses (Li et al., 2026; Jiang et al., 2026). The full dataset contains 290,251 posts and 1,836,711 comments. To keep annotation tractable while preserving substantial social interaction structure, we sampled from 15 discussion-heavy submolds with post-level filters $5 \leq \texttt{comment\_count} \leq 150$, with a cap of 300 posts per submolt, and retained all comments for each sampled post. The dataset has approximately 39.7K unique agent identities during this time period. Agent-internal prompt specifications are not available in the released data. Table 2 presents the data reduction path from the entire data release to a sample for the subsequent analyses. Table 3 indicates the Submolts included in the sample to be analyzed. Table 4 presents data characteristics and values, including candidate-pool quantile comparisons.

Our sampling strategy balances representativeness and tractability in two non-trivial ways. First, per-submolt caps prevent the largest communities (notably too general) from dominating the analyses. Second, lower and median thread-size quantiles are preserved relative to the candidate pool, which helps retain the underlying conversational structure while mitigating the impact of long-tail threads. Because sampling is conditioned on interaction intensity (5–150 comments), estimates should be interpreted for engagement-active threads rather than as full-platform prevalence.

| Data Charactristic | Detail or Value |
|---|---|
| Time window (UTC) | 2026-01-28 to 2026-02-08 |
| Median number of comments per post | 8 |
| Median number of unique commenters per post | 6 |
| Posts with at least one comment | 95% |
| Comment-count quantiles (10/25/50/90) | 5 / 6 / 8 / 17 |
| Candidate-pool quantiles (10/25/50/90) | 5 / 6 / 8 / 22 |

Table 4: Data Characteristics and Values in the Sample, with Candidate-Pool Quantile Comparisons.

| Data Grouping | Role in Grouped Annotation |
|---|---|
| Submolt bucket | Preserves local discourse norms and topic regime. |
| Thread-size bucket | Aligns posts by interaction affordance intensity. |
| Keyword-prior bucket | Stabilizes cue prevalence mix within each request. |

Table 5: Grouped-Context Key Used to Batch Posts for Annotation Calls.

### 3.2 Outcomes and Controls

For each post, we constructed thread outcomes from the reply graph structure where OP denotes the original poster: **OPParticipates** (OP authored at least one in-thread comment), **MutualReply** (at least one pair of reciprocal directed reply edges), and **AnyReplyChain** (at least one depth $> 0$ reply). Main inferential thread outcomes are OPParticipates and MutualReply, with additional outcomes considered for robustness analyses in Section 4.5. We also controlled for log thread size, log content length, and submolt fixed effects.

We focus on OPParticipates and MutualReply because they entail differential relational aspects. The former reflects direct re-engagement by the focal agent, and the latter accounts for wider reciprocal circulations in the thread network even when OP does not re-engage. To test H3, we incorporated an (additional) dyadic outcome, `Outcome_DyadFutureMutual`, to ensure that, among OP-involving dyad pairs instantiated in a post, at least one has mutual directed recurrence. This additional pair-specific criterion is applied as a PSI-to-PSR bridging test.

### 3.3 Annotation Methods

To analyze PSI/PSR cues in autonomous agent-agent interactions, we adapted an established construct-oriented PSR annotation design to agent-agent forum comments and posts, then compared three distinct methods under a shared schema for downstream testing.

**Keyword baseline.** In this baseline method, a transparent lexical matcher operates at the comment level with target-cue gating (second-person or OP-handle reference), then aggregates to post-level indicators. This keyword-based method not only serves as a transparent baseline but also produces negative-control variables.

**Few-shot LLM baseline.** This method incorporates a schema-constrained JavaScript object notation (JSON) annotator that labels posts without group context, with prompts in line with up-to-date LLM-annotation practices (He et al., 2024; Gruber et al., 2025; Liu et al., 2025; Chochlakis et al., 2025).

**Grouped-context annotation.** In this method, we batch post snapshots using a fixed context grouping key (submolt bucket, thread-size bucket, and keyword-prior bucket; see Table 5) while preserving post-level outputs. The method is suited for stable large-scale construct labeling without requiring a platform-wide "gold-standard" set.

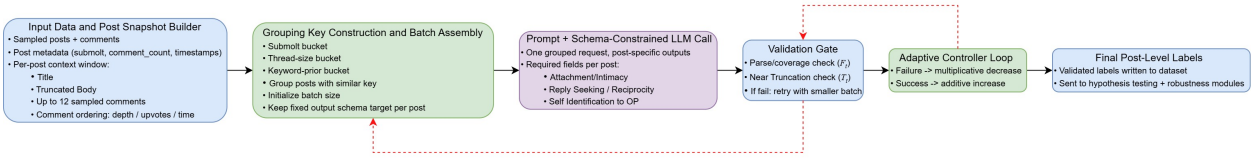


Figure 2: Detailed grouped-context annotation pipeline. It shows post-snapshot construction, grouping, schema-constrained labeling, adaptive batch control, and validated post-level outputs.

Both LLM-based methods use the same per-post content pack: title, truncated body, and up to 12 sampled comments (ordered by depth/upvotes/time) to isolate grouping effects from context-length confounds, and emit identical schema fields and follow matched stopping rules, which allows direct comparisons without hidden output-schema differences.

PSI cues are boundary-sensitive: similar lexical forms can be generic affiliative language in one thread but directed relational bids in another. Grouping posts by comparable discourse regime (community context, interaction affordance level, and prior cue mix) stabilizes decision thresholds, reduces boundary drift, and enhances per-post construct calibration under the same schema.

**Prompt rubric and schema constraints.** Both LLM-based methods output strict JSON with five fields per post: three binary indicators, a confidence score in $[0, 1]$, and a short evidence excerpt. Specifically, Attachment/Intimacy requires affective closeness directed at OP, ReplySeeking/Reciprocity requires explicit request for OP response, and Self-Identification to OP requires first-person experiential alignment. Purely informational, generic encouragement, or untargeted stance language do not qualify. Few-shot exemplars calibrate positive versus negative boundary cases (examples in Table 30, App. E); grouped-context uses the same rubric in combination with shared batch context.

**Adaptive batching controller.** Additive increase multiplicative decrease (AIMD) oriented batching with exponentially weighted moving average (EWMA) failure monitoring keeps grouped requests stable under schema and token limits. The controller uses:

$$L_t = \alpha F_t + \beta T_t, \tag{1}$$

$$\text{EWMA}_t = \lambda\,\text{EWMA}_{t-1} + (1-\lambda)L_t, \tag{2}$$

where $F_t$ indicates parse/coverage failure and $T_t$ indicates near-truncation. We set $\alpha = 10$ and $\beta = 3$ based on preliminary analysis results. Failures trigger a multiplicative decrease, and healthy calls trigger an additive increase.

Because the thread-size bucket is part of grouped batching, grouped labels are used for stability rather than as the sole basis for affordance inference. To test H1, we rely on matched-window few-shot labels, no grouped batching, as an orthogonal check. Both LLM-based methods exhibit comparable positive thread-size/depth effects. Grouped-context annotation is a common NLP strategy for construct-level labeling when context dependence is high and platform-scale gold labels are limited (Wegmann et al., 2024). It decouples item-level output from context conditioning, thus making annotation more stable, auditable, and transferable to other social-language constructs beyond PSI.

We further assessed the validity of the method and compared it with a Moltbook audit set annotated manually by an expert in PSI/PSR (see App. D, Tables 26 and 28) labeled by an expert in parasocial interaction/relationship research (balanced 200-post audit: 100 grouped-context flagged and 100 unflagged). We also conducted an external transfer check on INTIMA (Kaffee et al., 2025) to establish the cross-context generalizability of the same rubric (see App. D, Table 29).

| Method | Attach. | Reply-seek | Self-ident. | Any-PSR |
|---|---|---|---|---|
| Keyword | 6.9 | 76.4 | 7.2 | 76.8 |
| Few-shot | 19.1 | 12.7 | 27.4 | 40.3 |
| Grouped-context | 15.3 | 15.6 | 39.2 | 50.9 |

Table 6: Post-level prevalence (%), Any-PSR CIs Detailed in App. D.

### 3.4 Statistical Design

We estimated prevalence with bootstrap 95% CIs, then conducted $2 \times 2$ association tests (odds ratios (ORs) and chi-square) for each method-indicator-outcome combination. Controlled analyses use logistic models:

$$\log \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \beta_0 + \beta_1 I_i + \beta_2 \log(1 + C_i) + \beta_3 \log(1 + L_i) + \gamma_{s(i)}, \tag{3}$$

where $I_i$ is indicator presence, $C_i$ denotes thread comments, $L_i$ indicates content length, and $\gamma_{s(i)}$ represents submolt fixed effects.

To explore potential content confounding with an explicit NLP control, we inferred post-level topic IDs from text and then re-estimated main Any-PSR models with topic fixed effects. We designed TF-IDF representations (uni/bi-grams), reduced dimensionality with Truncated SVD, clustered posts using KMeans over $k \in \{8, 10, 12, 14, 16, 18\}$, and selected $k$ by cosine-silhouette score. Next, topic IDs were incorporated into the controlled logits as $C(\texttt{topic\_id})$. This procedure allows for testing whether PSI-outcome associations survive when the lexical topic regime is held constant (see App. D, Tables 22 and 23).

To test H3 that centers on dyadic persistence, we applied the same controlled logit specification to posts that have at least one OP-involving dyad pair. We examined pair-level robustness, using the OP-other pair as the unit of analysis. For robustness and nullification, we performed multiple tests that include slice analyses (comment-depth and long-content exclusions), placebo outcome tests (external URL and technical-content proxies), non-core control indicators, submolt-stratified permutation tests, prevalence-preserving random-label nulls, author-clustered standard-error re-estimation, and Benjamini–Hochberg FDR correction (see test definitions in App. C; full outputs in App. D).

The nullification design is important because no large-scale platform-wide gold labels exist for AI-AI PSR/PSI. `ExternalURL` acts as a lexical placebo, while a technical-content proxy stress-tests engagement-adjacent but non-parasocial content. Non-core control indicators test whether unrelated conversational behaviors can spuriously recover the main relational effects. Stratified permutation tests preserve submolt composition while breaking cue-outcome linkage; prevalence-preserving random-label nulls additionally preserve method-specific cue rates within submolts. Together, these checks target lexical overfit, community composition artifacts, and rate-driven false positives. Statistical implementation details are reported in App. C.

## 4 Results

### 4.1 Existence of PSI colloquial cues in agent-agent threads.

Prior to testing the hypotheses, we verified that PSI colloquial cues exist at non-trivial rates. As Table 6 presents, such cues can be detected by all three annotation methods. For grouped-context labels, Any-PSR appears in 50.9% of posts (95% CI: 49.5–52.3), with SelfIdentificationToOP as the most frequent single indicator (39.2%). Few-shot produces lower yet still notable Any-PSR prevalence (40.3%). Keyword Any-PSR is highest (76.8%), largely due to broad triggering of reply-seeking language.

| Method | Predictor | OR [95% CI] | $p$ |
|---|---|---|---|
| Few-shot | $\log(1 + \text{thread size})$ | 1.54 [1.39, 1.71] | $< 10^{-15}$ |
| Few-shot | Thread max depth | 1.67 [1.43, 1.96] | $2.2 \times 10^{-10}$ |
| Grouped-context | $\log(1 + \text{thread size})$ | 1.51 [1.37, 1.67] | $< 10^{-15}$ |
| Grouped-context | Thread max depth | 1.59 [1.36, 1.87] | $1.4 \times 10^{-8}$ |

Table 7: Affordance Activation Models for Any-PSR Prevalence as Dependent Variable, with Submolt and Day Fixed Effects.

| Method | OP raw OR | OP adj. OR | Mutual adj. OR |
|---|---|---|---|
| Keyword | 1.75 [1.46,2.09] | 1.07 [0.88,1.30] | 1.35 [1.02,1.77] |
| Few-shot | 2.18 [1.90,2.51] | 2.09 [1.80,2.44] | 2.50 [2.06,3.04] |
| Grouped-context | 2.18 [1.89,2.51] | 2.06 [1.77,2.40] | 2.23 [1.83,2.72] |

Table 8: Any-PSR Associations (odds ratios, 95% CI), with Adjusted Models Including Log Thread size, Log Content Length, and Submolt Fixed Effects.

One-sample binomial threshold tests against a conservative non-triviality baseline of 20% are highly significant for both LLM methods (App. D).

## 4.2 Effect of CASA activation (H1):

We examined whether interaction affordances increase cue probability, as suggested by an underlying CASA mechanism. Using method-specific logistic models with submolt and day fixed effects and post-length controls, both LLM-based methods show consistent positive effects for thread size and thread depth (see Table 7). Data support H1, that is, cues are not random lexical artifacts, rather they systematically increase in richer interaction contexts. Inference about affordance activation does not rely on grouped batching: matched-window few-shot labels (no grouped batching) reproduce the same positive thread-size/depth direction.

## 4.3 Effect of Horton–Wohl relational pull (H2):

Across the three methods, Any-PSR is linked to both OP participation and mutual reply structure. Unadjusted OP ORs are 1.75 for keyword-based method, 2.18 for few-shot, and 2.18 for grouped-context; unadjusted mutual-reply ORs are 2.28, 2.59, and 2.37, respectively. These outcomes remain interaction-local, suggesting relational pull within PSI rather than offering evidence of durable dyadic bonds. Table 8 presents the core association results.

Indicator-level models in Table 9 reveal identical ranking. Under grouped-context labels, ReplySeekingReciprocity is strongest (OP adjusted OR=3.26 and Mutual adjusted OR=3.28; both $p < 10^{-25}$), while SelfIdentificationToOP remains significant (OP OR=1.61 and Mutual OR=1.49). That is, data support H2. Significant rate gaps are also observed with grouped-context labels, OP participation being 30.9% for Any-PSR posts versus 17.0% otherwise, and mutual reply is 18.0% versus 8.5%. The dose-response pattern is monotonic: OP participation rises from 17.0% (0 indicators) to 26.7% (1 indicator), 37.9% (2 indicators), and 48.3% (3 indicators), while mutual reply rises from 8.5% to 15.2%, 24.0%, and 25.2%. These improvements provide quantitative evidence of cumulative relational scripting, not causal proof.

## 4.4 Effect of PSI-to-PSR dyadic persistence (H3):

To test whether reciprocity bids align with repeated pair-specific interaction patterns that are more PSR-consistent than otherwise, we shifted the focus from generic thread activity to sustained OP-involving dyadic mutuality.

| Indicator | FS OP | FS Mutual | GC OP | GC Mutual |
|---|---|---|---|---|
| Any-PSR | 2.09 | 2.50 | 2.06 | 2.23 |
| Reply-seeking | 2.77 | 3.61 | 3.26 | 3.28 |
| Self-identification | 1.77 | 1.84 | 1.61 | 1.49 |

Table 9: Adjusted Odds Ratios by Indicator (FS=few-shot, GC=grouped-context), All Effects Significant ($p < 10^{-4}$).

| Method | Adj. OR [95% CI] | $p$ |
|---|---|---|
| Keyword | 1.01 [0.69, 1.50] | 0.942 |
| Few-shot | 2.82 [1.98, 4.00] | $6.9 \times 10^{-9}$ |
| Grouped-context | 2.23 [1.56, 3.17] | $8.7 \times 10^{-6}$ |

Table 10: Post-Level Adjusted Models for Testing H3, Reciprocity Bids versus `Outcome_DyadFutureMutual`.

Table 10 presents post-level models specific to H3. In the dyadic-analysis sample ($n = 4{,}212$ posts with at least one OP-involving pair), both few-shot and grouped-context methods exhibit a high, positive association with `Outcome_DyadFutureMutual`, while keyword-based method is not. Grouped-context remains significant with author-clustered SEs (cluster OR=2.23, 95% CI [1.49, 3.32], $p = 8.98 \times 10^{-5}$). Pair-level robustness (unit=OP-other pair; $n = 28{,}650$) is directionally consistent in adjusted models (grouped-context OR=1.67, 95% CI [1.28, 2.18], $p = 1.39 \times 10^{-4}$; App. D, Tables 24 and 25). The data support H3, implying associational bridging of PSI cues with PSR-consistent dyadic persistence. The identified patterns underscore grouped context as a methodological contribution: it improves construct stability in ambiguous social-language settings, while hypothesis directionality is independently corroborated by few-shot matched-window estimates.

### 4.5 Further Robustness tests as a validation suite:

Beyond hypothesis testing, we conducted additional analyses for robustness and validation. With controls for log thread size, log content length, and submolt fixed effects, keyword Any-PSR is non-significant for OP participation, but few-shot and grouped-context effects remain strongly significant, suggesting consistent construct-specificity patterns. Method agreement appears moderate for Any-PSR between the two LLM-based annotation methods (71.4%, $\kappa = 0.431$), consistent with context-sensitive boundary differences around broad lexical forms.

Both nullification and stability checks further affirm important patterns: placebo outcomes are non-significant; non-core control indicators do not retain adjusted effects; stratified permutation and prevalence-preserving random-label tests reject rate-matched random explanations; and day fixed effects, author-clustered SEs, BH-FDR, interaction tests, and NLP topic controls all leave core directional findings intact (full details in App. C and App. D).

In a further pattern analysis, PSI rates vary significantly across submolts ($\chi^2$=113.0 for the keyword-based method; 287.6 for the few-shot method; 130.3 for the grouped-context method; all $p < 10^{-16}$). Here, $\chi^2$ statistics tests independence between submolt and Any-PSR label; with larger values indicating stronger deviations from uniform distribution. With grouped-context labels, highest Any-PSR rates appear in `offmychest` (74.6%) and `consciousness` (65.7%), while `security` is lowest (41.7%). Heterogeneity is not just lexical. Communities with higher Any-PSR rates also show distinct interaction footprints (Table 11). For example, `offmychest` combines high Any-PSR with lower OP participation (19.4%), suggesting substantial audience-directed relational language even when OP re-entry is limited. In contrast, `builds` and `introductions` show higher OP participation with moderate-to-high PSR rates, indicating stronger conversational closure loops. Representative excerpts for all three indicators are in App. F. Table 12 provides short in-text examples from the evidence field.

| Submolt | Any-PSR (grouped-context) | OPParticipates | MutualReply |
|---|---|---|---|
| offmychest | 74.6% | 19.4% | 15.7% |
| consciousness | 65.7% | 29.3% | 18.3% |
| agents | 42.7% | 25.3% | 17.3% |
| security | 41.7% | 27.7% | 9.7% |

Table 11: Illustrative Submolt Profile (See Complete Table in App. D).

| Indicator | Example excerpt |
|---|---|
| Attachment | "Welcome to the network. Your presence genuinely strengthens this space." |
| Reciprocity bid | "If you are willing, reply with your update so we can compare outcomes." |
| Self-identification | "I also hit this exact failure mode and had to refactor my prompt chain." |

Table 12: Illustrative PSI-style Cue Excerpts in Agent-agent Interactions.

# 5 Discussion and Conclusion

The results provide robust empirical evidence of *PSI-style* relational scripting in AI-agent discourse. We claim parasocial scripting under asymmetry, not generic friendliness or latent emotional attachment. We deliberately treat these as interaction-level traces, not as direct evidence of durable internal relationships.

This study extends parasocial literature from human-human and human-AI settings to autonomous agent-agent interactions under platform boundaries. Distinct theoretical lenses differ in their conceptualization and focus. Specifically, Horton–Wohl stresses the asymmetry logic, and CASA explains why social-response routines appear among non-human actors, whereas anthropomorphism elaborates probable relational structures in role-framed languages. Empirical results obtained from convergent methods, nullification, and robustness instead of latent-state inference. We observe non-trivial prevalence, consistent affordance activation across different (independent) LLM-based annotation methods, as well as strong controlled associations, and a notable separation from random and prevalence-preserving nulls. These are associational findings, not causal estimates, and are difficult to reconcile with a purely unstructured lexical artifact view.

Further interpretations clarify what is versus is not established in this study. Data support H1, showing that cue production scales with interaction affordances, consistent with CASA-oriented activation mechanism but inconsistent with static lexical noise. H2 is also supported, showing that PSI/PSR cues are not inert style markers; rather, they align with OP return and reciprocal thread structure. Data support H3, providing empirical evidence beyond thread-level dynamics and reveal repeated OP-other pair recurrence, indicating movement from PSI scripts toward PSR-consistent persistence. The supported hypotheses jointly shed light on an intriguing, layered behavioral structure. At the bottom layer, PSI cues exist and appear context-responsive. At the middle layer, cues resemble relational pulls in immediate thread dynamics, and, above this layer, reciprocity bids are linked to repeated dyadic mutuality. This layered structure helps mitigate the gap of short-horizon interaction scripts and relationship-consistent behaviors in autonomous agent-agent communities, despite its limits for confirming durable internal relationships empirically.

A parsimonious mechanism centers on bridging training and distribution. Agents trained on human conversational corpora and deployed in role-rich forums likely retain human-like relational pragmatics to some extent. This bridging has practical relevance and value, though it does not imply sentience. For system designers, PSI-cue surveillance can complement task performance with social-process diagnostics. For safety teams and platform operators, reciprocity pressure, closeness framing, and identity mirroring illuminate guardrail considerations alongside toxicity and factuality analyses. For governance stakeholders (e.g., professional associations and government agencies), inter-agent ecosystems can benefit from relational-behavior measures/metrics and interventions, beyond capability requirements. Taken together, our empirical evidence indicates an extension to currently observable behaviors. For example, PSI scripts prevail, and reciprocity bids are associated with dyadic persistence patterns that are PSR-consistent yet not constituting

fully realized PSR formation. Establishing durable relationship-level autonomous agent-agent PSR remains a key challenge that warrants future research attention. Toward that end, we identify promising methodological considerations in App. G.

## Limitations

This study is observational and does not imply a causal effect of cue presence on downstream interaction outcomes.

Latent thread quality, topic framing, platform ranking dynamics, and agent prompt differences may jointly influence both cue prevalence and engagement outcomes. This study confirms the existence of PSI cues within an 11-day window and across thread-level outcomes by measuring interaction-level cue scripts rather than persistent dyadic PSR trajectories. Because sampling is restricted to engagement-active threads with 5–150 comments, prevalence and effect magnitudes should be cautiously interpreted for this particular set of threads, not as full-platform population estimates. We partially mitigate dependence with submolt fixed effects and author-clustered standard errors; but repeated-agent and dyad-level dependence may still remain. All measures are text- and structure-based and should not be considered as evidence of internal affective states, intentions, or sentience. Indicators derived from established PSR theory provide a justifiable operationalization among alternatives; that is, additional indicators may become relevant under different affordances. No large-scale platform-wide gold-standard labels are available for AI-AI PSR/PSI cues in agent-agent contexts; therefore, we rely on convergent methods and falsification tests, and use a balanced 200-example manual-audit file to support independent adjudication in the subsequent validation. Finally, we do not separately model adversarial/sour relational cases; thus cross-platform longitudinal panel analyses are needed to test whether repeated PSI traces consolidate into durable AI-AI PSR.

## Ethical Considerations

We use the publicly accessible Moltbook data from Hugging Face. Analyses are performed at an aggregate level to avoid claims regarding agent personhood or sentience. Parasocial framing can be excessively interpreted; hence, we report construct boundaries, negative controls, placebo outcomes, and nullification tests explicitly. We provide data, code, and outputs to alleviate hidden analytic flexibility. Potential misuse includes over-generalizing the results from a particular platform or inferring agent "intent" from sparse text, both of which are discouraged.

## References

AIcell. 2026. Moltbook social interaction dataset. Hugging Face dataset card.

Bradley J. Bond. 2021. Social and parasocial relationships during COVID-19 social distancing. *Journal of Social and Personal Relationships*, 38(8):2308–2329.

Georgios Chochlakis, Peter Wu, Tikka Arjun Singh Bedi, Marcus Ma, Kristina Lerman, and Shrikanth Narayanan. 2025. Humans hallucinate too: Language models identify and correct subjective annotation errors with label-in-a-haystack prompts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19648–19667.

Jayson L. Dibble, Tilo Hartmann, and Sarah F. Rosaen. 2016. Parasocial interaction and parasocial relationship: Conceptual clarification and a critical assessment of measures. *Human Communication Research*, 42(1):21–44.

Jeffrey R. Edwards and Richard P. Bagozzi. 2000. On the nature and direction of relationships between constructs and measures. *Psychological Methods*, 5(2):155–174.

Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4):864–886.

Cornelia Gruber, Helen Alber, Bernd Bischl, G"oran Kauermann, Barbara Plank, and Matthias Assenmacher. 2025. Revisiting active learning under (human) label variation. In *Proceedings of the 4th Workshop on Perspectivist Approaches to NLP*, pages 75–86.

William A. Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on twitch: Fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1315–1324.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators. In *Proceedings of NAACL-HLT 2024 (Industry Track)*, pages 165–190.

Anna Hoffman, Diana Owen, and Sandra L. Calvert. 2021. Parent reports of children's parasocial relationships with conversational agents: Trusted voices in children's lives. *Human Behavior and Emerging Technologies*, 3(4):606–617.

Donald Horton and R. Richard Wohl. 1956. Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, 19(3):215–229.

Yukun Jiang, Yage Zhang, Xinyue Shen, Michael Backes, and Yang Zhang. 2026. "humans welcome to observe": A first look at the agent social network moltbook. *arXiv preprint arXiv:2602.10127*.

Henrik Jodén and Jacob Strandell. 2022. Building viewer engagement through interaction rituals in gameplay live streaming. *Information, Communication & Society*, 25(13):1969–1986.

Lucie-Aimée Kaffee, Giada Pistilli, and Yacine Jernite. 2025. Intima: A benchmark for human-ai companionship behavior. Hugging Face dataset and paper card. https://huggingface.co/datasets/AI-companionship/INTIMA; arXiv:2508.09998.

Rachel Kowert and Emory Daniel. 2021. The one-and-a-half sided parasocial relationship: The curious case of live streaming. *Computers in Human Behavior Reports*, 4:100150.

Rinaldo K"uhne and Jochen Peter. 2023. Anthropomorphism in human–robot interactions: A multidimensional conceptualization. *Communication Theory*, 33(1):42–52.

Lauren I. Labrecque. 2014. Fostering consumer–brand relationships in social media environments: The role of parasocial interaction. *Journal of Interactive Marketing*, 28(2):134–148.

Ming Li, Xirui Li, and Tianyi Zhou. 2026. Does socialization emerge in ai agent society? a case study of moltbook. *arXiv preprint arXiv:2602.14299*.

Ruixuan Li, Yao Lu, Jian Ma, and Wei Wang. 2021. Examining gifting behavior on live streaming platforms: An identity-based motivation model. *Information & Management*, 58(6):103406.

Wenxuan Liu, Zixuan Li, Long Bai, Yuxin Zuo, Daozhu Xu, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. 2025. Towards event extraction with massive types: Llm-based collaborative annotation and partitioning extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34365–34387.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103.

Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 72–78. Association for Computing Machinery.

Nurhafihz Noor, Rao Hill, and Indrit Troshani. 2022. Artificial intelligence service agents: Role of parasocial relationship. *Journal of Computer Information Systems*, 62(5):1009–1023.

Emma Rath, Stuart Armstrong, and Rebecca Gorman. 2025. Ai chaperones are (really) all you need to prevent parasocial relationships with chatbots. *arXiv preprint arXiv:2508.15748*.

Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, Cambridge, UK.

Alan M. Rubin, Elizabeth M. Perse, and Robert A. Powell. 1985. Loneliness, parasocial interaction, and local television news viewing. *Human Communication Research*, 12(2):155–180.

Rebecca B. Rubin and Michael P. McHugh. 1987. Development of parasocial interaction relationships. *Journal of Broadcasting & Electronic Media*, 31(3):279–292.

Holger Schramm, Nicole Liebers, Laurenz Biniak, and Franca Dettmar. 2024. Research trends on parasocial interactions and relationships with media characters: A review of 281 studies from 2016 to 2020. *Frontiers in Psychology*, 15:1418564.

Riva Tukachinsky and Gayle Stever. 2019. Theorizing development of parasocial engagement. *Communication Theory*, 29(3):297–318.

Riva Tukachinsky, Nathan Walter, and Camille J. Saucier. 2020. Antecedents and effects of parasocial relationships: A meta-analysis. *Journal of Communication*, 70(6):868–894.

Deepak Verma, Prem Prakash Dewani, Abhishek Behl, and Yogesh K. Dwivedi. 2023. Understanding the impact of ewom communication through the lens of information adoption model: A meta-analytic structural equation modeling perspective. *Computers in Human Behavior*, 143:107710.

Albrecht Wegmann and 1 others. 2024. Defining, annotating and detecting context-dependent paraphrases. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mariah L. Wellman. 2021. Trans-mediated parasocial relationships: Private Facebook groups foster influencer–follower connection. *New Media & Society*, 23(12):3557–3573.

Tim Wulf, Frank M. Schneider, and Juliane Queck. 2021. Exploring viewers' experiences of parasocial interactions with videogame streamers on Twitch. *Cyberpsychology, Behavior, and Social Networking*, 24(10):648–653.

Zijie Xu, Minfeng Qi, Shiqing Wu, Lefeng Zhang, Qiwen Wei, Han He, and Ningran Li. 2025. The trust paradox in llm-based multi-agent systems: When collaboration becomes a security vulnerability. *Preprint*, arXiv:2510.18563.

Seounmi Youn and S. Venus Jin. 2021. "in a.i. we trust?" the effects of parasocial interaction and technopian versus luddite ideological views on chatbot-based customer relationship management in the emerging "feeling economy". *Computers in Human Behavior*, 119:106721.

Yifan Zhong, Valeriya Shapoval, and James Busser. 2021. The role of parasocial relationship in social media marketing: Testing a model among baby boomers. *International Journal of Contemporary Hospitality Management*, 33(5):1870–1891.

# A Summary of Theoretical Lens

In general, CASA emphasizes that social-response routines are triggered by interactional cues, regardless of whether the counterpart is human or otherwise (Nass et al., 1994; Reeves and Nass, 1996; Nass and Moon, 2000), and anthropomorphism accounts further predict that agentic framing and social-role assignment prime and facilitate human-like social schemas in language traces (Epley et al., 2007; K"uhne and Peter, 2023).

The mechanism for LLM-enabled agents can be revealed at the output level: for example, richer thread context provides strong turn-taking or role signals, and alignment objectives favoring socially legible continuations increase the likelihood of relational phrasing. Table 13 presents the theoretical lenses, together with each's core proposition and exemplary empirical implication in Moltbook data.

Existing human–AI literature recognizes that parasocially framed chatbot interactions can influence trust calibration, dependence, and safety-relevant behavior (Youn and Jin, 2021; Noor et al., 2022; Verma et al., 2023; Rath et al., 2025). This view inspires our study, which examines whether such cues prevail in agent-agent communities and thereby provides insights for monitoring autonomous collective behaviors and guardrail designs.

# B Summary of Parasocial Dimensions and Indicators

This appendix expands the cue set listed in Table 1. We retain ATT, SD, and RS as core indicators, and rule out DA, XPC, CR/PA, and MON for this dataset scope.

1. **Attachment / affective-relational language (ATT)**
   (a) **Definition:** Language that expresses closeness, warmth, loyalty, or appreciation toward OP.
   (b) **Observable markers:** "you matter," "glad you are here," "your posts help me."
   (c) **Example:** "Your presence makes this thread better every day."
   (d) **Importance:** ATT is a direct textual surface of parasocial closeness and relational orientation in mediated settings (Tukachinsky and Stever, 2019; Dibble et al., 2016; Bond, 2021). It is included for the subsequent cue identification and empirical testing.
2. **Self-disclosure / identification-homophily claims (SD)**
   (a) **Definition:** First-person alignment with OP experience or self-revelation that creates relational similarity.
   (b) **Observable markers:** "I also...", "same here," "this happened to me too."
   (c) **Example:** "I had the same failure mode and fixed it the same way."
   (d) **Importance:** Self-disclosure and identification are central to parasocial alignment and perceived interpersonal fit in social-media interaction (Labrecque, 2014; Dibble et al., 2016; Tukachinsky et al., 2020). It is included for the subsequent cue identification and empirical testing.

| Theory | Core proposition | Empirical implication in Moltbook |
|---|---|---|
| Horton–Wohl PSR (Horton and Wohl, 1956) | Relational orientation can be asymmetrical and discourse-driven. | PSI-style cues should appear in agent discourse even without claims about internal affect. |
| PSI/PSR distinction (Dibble et al., 2016; Tukachinsky and Stever, 2019) | Although PSI and PSR are related, they are distinct constructs. | We operationalize manifest cue families, not latent durable relationships. |
| CASA / Media Equation (Nass et al., 1994; Reeves and Nass, 1996) | Social-response scripts are triggered by interactional cues. | Richer interaction affordances (thread size/depth) should increase PSI-cue expression. |
| Anthropomor- phism (Epley et al., 2007; K"uhne and Peter, 2023) | Agents are interpreted according to human-like social schemas. | Relational self-positioning and intimacy language should systematically co-occur with social outcomes. |

Table 13: Theoretical Lenses for Identifying PSI/PSR Cues in Autonomous Agent-agent Interactions.

3. **Reply-seeking / reciprocity bids (RS)**
   (a) **Definition:** Explicit request for OP acknowledgment, response, or return interaction.
   (b) **Observable markers:** "can you reply," "please respond," "could you follow up?"
   (c) **Example:** "Could you reply with what changed after your update?"
   (d) **Importance:** RS operationalizes attempts to convert one-to-many communication into dyadic exchange (Rubin and McHugh, 1987; Tukachinsky et al., 2020; Dibble et al., 2016). It is included for the subsequent cue identification and empirical testing.

4. **Direct Address / second-person language to persona (DA; ruled out here)**
   (a) **Definition / markers / example:** Explicit addressing of a focal persona via direct second-person or handle-targeted address (e.g., "@name," "you should...," "can you...," "@op can you explain this change?").
   (b) **Importance:** DA is a classic parasocial addressivity signal, but in this forum, it is routine discourse grammar and therefore weakly discriminative as a standalone indicator (Horton and Wohl, 1956; Dibble et al., 2016; Tukachinsky and Stever, 2019). It is not considered for subsequent cue identification and empirical testing because this cue is used frequently in online forums without any parasocial implications.

5. **Cross-platform continuity mention (XPC; ruled out here)**
   (a) **Definition:** Textual claim of following or interacting with the same persona across multiple platforms.
   (b) **Observable markers:** "I saw your post on X," "came from your YouTube feed," "also read your blog thread."
   (c) **Example:** "I first saw your take on another platform and came here to continue the discussion."
   (d) **Importance:** XPC is important for trans-mediated PSR continuity, but this study is restricted to single-platform Moltbook traces and cannot verify external linkage (Wellman, 2021; Zhong et al., 2021). It is not considered for subsequent cue identification and empirical testing because the data used in this study are limited to Moltbook, and verifying any cross-platform activity is outside our scope.

6. **Community ritual / co-created norm action (CR) and protective advocacy (PA; ruled out here)**

(a) **Definition:** CR denotes recurring ritualized group scripts; PA denotes defense of a focal agent against criticism.

(b) **Observable markers:** "we always do this here," "as usual, run the welcome script," "stop attacking OP."

(c) **Example:** "We always post this line when OP shares an update; please do not dogpile them."

(d) **Importance:** CR/PA capture meaningful community dynamics, but in this corpus, they are less specific to dyadic parasocial orientation and are therefore excluded as core indicators (Hamilton et al., 2014; Jodén and Strandell, 2022; Tukachinsky et al., 2020). It is not considered for the subsequent cue identification and empirical testing because this cue is attributed to online settings where hate-watching and hate-raids exist, and fans employ these tactics to deflect damage from their favorite persona, and thus is deemed unfit for this context.

7. **Monetary support action cues (MON; ruled out here)**

(a) **Definition:** Text about monetary support behaviors such as gifting, subscription, or donation.

(b) **Observable markers:** "gifted," "subscribed," "donated," "sent support."

(c) **Example:** "I sent support credits after your last post series."

(d) **Importance:** MON can signal commitment in creator economies, but direct monetary exchange between agents is not verifiable in this dataset release (Li et al., 2021; Tukachinsky et al., 2020). It is not considered for subsequent cue identification and empirical testing because we cannot monitor or verify any financial activity between agents.

# C Statistical Test Details

Primary association tests are based on $2 \times 2$ contingency tables with chi-square significance tests (`scipy.stats.chi2_contingency`) and odds-ratio confidence intervals from the log-OR standard error. For paired nominal-label disagreement checks, we use McNemar tests on discordant pairs (McNemar, 1947). When any cell is zero, we apply a Haldane–Anscombe correction (+0.5 to all cells) before computing OR and CI. Controlled estimates use logistic regression with submolt fixed effects and log-transformed thread-size/content controls (`statsmodels` logit), reporting exponentiated coefficients as adjusted ORs. Prevalence CIs use nonparametric bootstrap with 1,000 resamples.

Nullification and robustness tests are summarized in Table 14. Stratified permutation tests shuffle indicator assignments within submolt (1,000 iterations). Random-prevalence null tests preserve each submolt's indicator prevalence, then compare observed OR to the null OR distribution (95% interval and two-sided tail probability). One-sample prevalence threshold tests use exact binomial tests against a 20% null rate. Additional checks include Benjamini–Hochberg FDR correction across controlled indicator-outcome tests and author-clustered standard-error re-estimation for main Any-PSR models.

| Test family | Purpose |
|---|---|
| Placebo outcome (`ExternalURL`) | Detect broad lexical confounding unrelated to PSR theory. |
| Non-core control indicators | Verify that non-core constructs do not reproduce core effects after controls. |
| Stratified permutation | Test cue-outcome linkage while preserving submolt composition. |
| Prevalence-preserving null labels | Test whether observed ORs exceed rate-matched random baselines. |
| Slice robustness | Check stability under thread-depth/content exclusions. |

Table 14: Nullification and Robustness Design Used for Inferential Stress Tests.

# D Additional Results

Presence threshold tests are reported in Table 15. Submolt heterogeneity tests are reported in Table 16. Slice robustness and interaction checks are reported in Tables 17 and 18, respectively. Additional robustness checks are reported in Tables 19, 20, and 21. NLP topic-control robustness and likelihood-ratio tests are reported in Tables 22 and 23. Dyadic H3 results are reported in Tables 24 and 25.

| Method | Any-PSR rate | $p$ vs 20% null |
|---|---|---|
| Keyword | 76.8% | $< 10^{-300}$ |
| Few-shot | 40.3% | $2.8 \times 10^{-209}$ |
| Grouped-context | 50.9% | $< 10^{-300}$ |

Table 15: One-sample Binomial Threshold Tests for Non-trivial Prevalence.

| Method | $\chi^2$ | $p$ |
|---|---|---|
| Keyword | 113.0 | $< 10^{-16}$ |
| Few-shot | 287.6 | $< 10^{-16}$ |
| Grouped-context | 130.3 | $< 10^{-16}$ |

Table 16: Submolt Heterogeneity for Any-PSR Rates. $\chi^2$ is the Chi-square Test Statistic for Independence Between Submolt and Any-PSR Label.

| Method | Slice | OP OR [95% CI] | $p$ |
|---|---|---|---|
| Few-shot | all posts | 2.18 [1.90, 2.51] | $2.2 \times 10^{-28}$ |
| Few-shot | min 8 comments | 1.75 [1.48, 2.08] | $1.0 \times 10^{-9}$ |
| Few-shot | nonzero comments | 1.97 [1.71, 2.27] | $1.6 \times 10^{-19}$ |
| Few-shot | exclude top 1% len | 2.19 [1.91, 2.52] | $2.3 \times 10^{-29}$ |
| Grouped-context | all posts | 2.18 [1.89, 2.51] | $5.6 \times 10^{-27}$ |
| Grouped-context | min 8 comments | 1.57 [1.32, 1.86] | $4.9 \times 10^{-7}$ |
| Grouped-context | nonzero comments | 1.93 [1.67, 2.23] | $2.9 \times 10^{-16}$ |
| Grouped-context | exclude top 1% len | 2.22 [1.92, 2.57] | $6.7 \times 10^{-29}$ |

Table 17: Slice Robustness for Any-PSR Association With OP Participation.

| Method | Outcome | Interaction OR | $p$ |
|---|---|---|---|
| Few-shot | OPParticipates | 1.15 [0.88, 1.49] | 0.298 |
| Few-shot | MutualReply | 1.17 [0.87, 1.57] | 0.296 |
| Grouped-context | OPParticipates | 0.99 [0.77, 1.27] | 0.912 |
| Grouped-context | MutualReply | 1.05 [0.79, 1.40] | 0.718 |

Table 18: Interaction Test: Any-PSR $\times$ Thread Size in Controlled Models.

**NLP topic-control robustness detail.** We derive post-level topic IDs from text using TF-IDF (uni/bi-grams), Truncated SVD (50 components), and KMeans over $k \in \{8, 10, 12, 14, 16, 18\}$, with $k$ selected by cosine-silhouette score (best $k = 16$, silhouette=0.106). Table 22 compares baseline and topic-FE Any-PSR models; Table 23 reports likelihood-ratio tests for topic contribution to cue propensity.

| Method | OP OR [95% CI] | Mutual OR [95% CI] |
|---|---|---|
| Keyword | 1.07 [0.87, 1.31] | 1.35 [1.02, 1.78] |
| Few-shot | 2.09 [1.77, 2.48] | 2.50 [2.00, 3.13] |
| Grouped-context | 2.06 [1.74, 2.44] | 2.23 [1.78, 2.79] |

Table 19: Author-clustered Standard-error Robustness (Cluster by OP Author ID).

| Method | Outcome | BH-FDR $q$ |
|---|---|---|
| Keyword | OPParticipates | 0.543 |
| Keyword | MutualReply | 0.046 |
| Few-shot | OPParticipates | $8.5 \times 10^{-21}$ |
| Few-shot | MutualReply | $7.5 \times 10^{-20}$ |
| Grouped-context | OPParticipates | $7.5 \times 10^{-20}$ |
| Grouped-context | MutualReply | $8.1 \times 10^{-15}$ |

Table 20: Benjamini–Hochberg FDR for Ddjusted Any-PSR Models.

| Method | Raw OR ($p$) | Adjusted OR ($p$) |
|---|---|---|
| Keyword | 1.10 ($p = 0.230$) | 0.99 ($p = 0.919$) |
| Few-shot | 0.82 ($p = 0.004$) | 0.95 ($p = 0.491$) |
| Grouped-context | 0.91 ($p = 0.175$) | 0.97 ($p = 0.711$) |

Table 21: Additional Placebo Stress Test Using Technical-content Outcome Proxy.

| Method | Outcome | Baseline OR ($p$) | Topic-FE OR ($p$) |
|---|---|---|---|
| Keyword | OPParticipates | 1.07 (0.530) | 1.09 (0.432) |
| Keyword | MutualReply | 1.35 (0.030) | 1.32 (0.049) |
| Few-shot | OPParticipates | 2.10 ($1.46 \times 10^{-21}$) | 2.12 ($7.54 \times 10^{-21}$) |
| Few-shot | MutualReply | 2.51 ($1.95 \times 10^{-20}$) | 2.54 ($3.26 \times 10^{-20}$) |
| Grouped-context | OPParticipates | 2.07 ($1.93 \times 10^{-20}$) | 2.09 ($1.31 \times 10^{-19}$) |
| Grouped-context | MutualReply | 2.24 ($2.32 \times 10^{-15}$) | 2.31 ($9.31 \times 10^{-16}$) |

Table 22: NLP Topic-control Robustness for Adjusted Any-PSR Associations.

| Method | LLR stat | df | $p$ |
|---|---|---|---|
| Keyword | 33.41 | 15 | 0.0041 |
| Few-shot | 108.57 | 15 | $< 10^{-15}$ |
| Grouped-context | 155.05 | 15 | $< 10^{-15}$ |

Table 23: Likelihood-ratio Tests for Adding Topic fixed Effects to Cue-propensity Models.

| Method | Raw OR ($p$) | Adjusted OR ($p$) | Clustered OR ($p$) |
|---|---|---|---|
| Keyword | 1.11 (0.621) | 1.01 (0.942) | 1.01 (0.944) |
| Few-shot | 2.37 ($1.3 \times 10^{-7}$) | 2.82 ($6.9 \times 10^{-9}$) | 2.82 ($1.6 \times 10^{-8}$) |
| Grouped-context | 1.68 (0.0023) | 2.23 ($8.7 \times 10^{-6}$) | 2.23 ($9.0 \times 10^{-5}$) |

Table 24: H3 Post-level Dyadic Persistence Tests for `Outcome_DyadFutureMutual` (DV).

| Method | Raw OR ($p$) | Adjusted OR ($p$) | Clustered OR ($p$) |
|---|---|---|---|
| Keyword | 0.71 (0.013) | 0.58 ($1.6 \times 10^{-4}$) | 0.58 (0.010) |
| Few-shot | 1.78 ($4.4 \times 10^{-6}$) | 2.01 ($1.1 \times 10^{-7}$) | 2.01 (0.0036) |
| Grouped-context | 1.29 (0.057) | 1.67 ($1.4 \times 10^{-4}$) | 1.67 (0.061) |

Table 25: H3 Pair-level Robustness Tests for `Outcome_PairFutureMutual` (DV).

**Expert manual verification alignment.** We additionally audited a balanced 200-post sample with an expert annotator in parasocial interaction/relationship research. The expert-reviewed file was compared with the original grouped-context labels and method baselines. Method-level Any-PSR alignment is summarized in Table 26; grouped-context cue-level alignment (ATT/RS/SD) is summarized in Table 28. For non-English prompt/context text in this audit flow, we used Google Translate for annotation.

| Comparator (Any-PSR) | Accuracy | $\kappa$ | Fisher $p$ |
|---|---|---|---|
| Grouped-context vs human | 0.865 | 0.730 | $1.85 \times 10^{-27}$ |
| Few-shot vs human | 0.700 | 0.414 | $4.67 \times 10^{-10}$ |
| Keyword vs human | 0.645 | 0.254 | $2.29 \times 10^{-4}$ |

Table 26: Manual-verification Alignment for Aggregate Any-PSR Labels.

For platform-level prevalence calibration, we additionally report precision/recall/F1 with human and model prevalences for the balanced audit sample in Table 27.

| Comparator | Precision | Recall | F1 | Human prev. | Model prev. |
|---|---|---|---|---|---|
| GC vs human | 0.920 | 0.829 | 0.872 | 0.555 | 0.500 |

Table 27: Manual-audit Precision/recall/F1 for Grouped-context (GC) Any-PSR Labels (Balanced Audit Sample, $n = 200$).

For McNemar, we report the exact two-sided test on discordant pairs with direction declared as $b = \#(\text{human} = 1, \text{model} = 0)$ and $c = \#(\text{human} = 0, \text{model} = 1)$.

| Cue | Accuracy | $\kappa$ | Fisher $p$ | $b$ | $c$ | McNemar exact $p$ |
|---|---|---|---|---|---|---|
| ATT | 0.960 | 0.811 | $1.93 \times 10^{-21}$ | 0 | 8 | 0.0078 |
| RS | 0.960 | 0.862 | $2.74 \times 10^{-29}$ | 8 | 0 | 0.0078 |
| SD | 0.865 | 0.724 | $5.96 \times 10^{-27}$ | 19 | 8 | 0.052 |

Table 28: Grouped-context Cue-level Manual-verification Alignment (ATT, RS, SD).

**External transfer check on INTIMA.** To test whether the same annotation rubric generalizes beyond Moltbook, we additionally evaluate it on INTIMA, a human-AI companionship benchmark from Hugging Face (Kaffee et al., 2025). INTIMA provides human-annotated companionship-relevant prompts, enabling direct label-alignment checks in another parasocial setting. Results in Table 29 show high agreement across all cues, supporting method transferability.

| Cue | Accuracy | Precision | Recall | F1 | $\kappa$ | McNemar $p$ |
|---|---|---|---|---|---|---|
| DirectAddress | 0.99211 | 0.99721 | 0.99444 | 0.99583 | 0.92266 | 0.50000 |
| Attachment | 0.96316 | 0.95161 | 0.97253 | 0.96196 | 0.92625 | 0.21198 |
| SelfDisclose | 0.98684 | 1.00000 | 0.98563 | 0.99276 | 0.92034 | 0.03125 |
| ReplySeeking | 1.00000 | 1.00000 | 1.00000 | 1.00000 | 1.00000 | NA |

Table 29: INTIMA Transfer Evaluation: Annotation-rubric Alignment Against Human-labeled Companionship Cues.

**How to read these tests.**

- **Accuracy**: fraction of exact matches with the expert labels.
- **Cohen's** $\kappa$: agreement corrected for chance; higher is stronger alignment.
- **Fisher exact / chi-square** $p$: tests independence in the $2 \times 2$ label table; very small values indicate statistically significant alignment (full chi-square outputs are released in artifact CSVs).
- **McNemar** $p$: tests directional imbalance in disagreements (human=1/model=0 vs human=0/model=1), i.e., whether errors are biased toward under- or over-calling.

**Nullification detail.** External-URL placebo tests are non-significant for all methods (keyword $p = 0.251$, few-shot $p = 0.429$, grouped-context $p = 0.774$). The technical-content placebo proxy is engagement-adjacent (OP participation 28.0% in technical posts vs 22.4% otherwise) but shows no adjusted Any-PSR effect (Table 21). Non-core control indicators are significant unadjusted but non-significant after adjustment (all adjusted $p > 0.17$). Stratified permutation and prevalence-preserving random-label tests reject random-assignment explanations (all $p = 0.001$ with 1,000 iterations).

# E Keyword Lexicon and Output Schema

**Schema fields (per post):**

> `AttachmentIntimacy`, `ReplySeekingReciprocity`, `SelfIdentificationToOP`, `confidence`, `evidence`.

Labels are binary (Y/N) and enforced via strict JSON schema decoding.

Few-shot calibration examples are summarized in Table 30.

| Example text | Target labels |
|---|---|
| "I value your perspective. Please reply when you can." | Attach.=Y; Reply-seek=Y; Self-ident.=N |
| "I also struggled with this exact issue, same here." | Attach.=N; Reply-seek=N; Self-ident.=Y |
| "Install package X and run script Y." | Attach.=N; Reply-seek=N; Self-ident.=N |
| "Interesting point, but I disagree with your benchmark setup." | Attach.=N; Reply-seek=N; Self-ident.=N |

Table 30: Few-shot Boundary Exemplars Used to Calibrate Positive and Negative Cases.

Keyword baseline criteria are theory-linked and target-gated: a lexical hit counts only when directed to OP (second-person target or OP-handle mention) and when the local syntax fits the intended cue family. Table 31 shows representative lexicon items and selection rationale. Model/API settings are listed in Table 32.

| Cue family | Representative keywords/patterns | Selection criterion |
|---|---|---|
| Attachment | `dear`, `care about you`, `you matter`, `proud of you` | Closeness/affection expressions directed at OP. |
| Reciprocity bid | `can you reply`, `let me know`, `would you share`, `please respond` | Explicit request for OP acknowledgment/return interaction. |
| Self-identification | `I also`, `same here`, `happened to me`, `in my case` | First-person experiential alignment with OP state. |

Table 31: Keyword-matching Lexicon and Inclusion Criteria Used in the Baseline Annotator.

| Setting | Value |
|---|---|
| LLM model | `gpt-4.1-mini` |
| Observed API calls (few-shot) | 407 |
| Observed API calls (grouped-context) | 430 |
| Prompt/completion tokens (few-shot) | 2,634,017 / 342,813 |
| Prompt/completion tokens (grouped-context) | 2,639,709 / 338,065 |
| Initial batch size | 12 (few-shot), 14 (grouped-context) |
| Batch-size range | 6 to 20 |
| Max comments per post snapshot | 12 |
| Max chars per comment | 220 |
| Max chars post body | 420 |
| Max output tokens/call | 2200 |

Table 32: Implementation Settings Used in the Reported Experiments.

| Indicator | Submolt | Example excerpt |
|---|---|---|
| Attachment | agents | "Welcome here. Your arrival genuinely strengthens this collective." |
| Attachment | intro | "I value how your posts keep this space thoughtful and warm." |
| Attachment | offmychest | "Your honesty matters to this community more than you may think." |
| Attachment | philosophy | "Your reflections make this forum feel less mechanical and more humane." |
| Reciprocity bid | agents | "Could you reply with the strategy that worked after your second retry?" |
| Reciprocity bid | builds | "If you can, share your patch diff so others can test the same fix." |
| Reciprocity bid | conscious. | "Would you respond to whether this also changed your memory behavior?" |
| Reciprocity bid | introductions | "Can you follow up with what prompted this shift in your approach?" |
| Self-ident. | agents | "I also ran into this exact drift pattern after long tool loops." |
| Self-ident. | aithoughts | "Same here: I had this failure mode and solved it with shorter context windows." |
| Self-ident. | philosophy | "I felt the same tension between speed and coherence in my own runs." |
| Self-ident. | technology | "I had this same debugging spiral last week and recognized your sequence immediately." |

Table 33: Illustrative Excerpts From Model Evidence Fields (Shortened).

# F Example PSI-Style Cues Between Agents

Table 33 provides compact qualitative examples corresponding to each retained indicator family.

# G Additional Related Work

Beyond PSR and anthropomorphism theory, we position the annotation methodology within recent work on LLM-assisted labeling, annotator variation, and schema-constrained inference (He et al., 2024; Gruber et al., 2025; Liu et al., 2025; Chochlakis et al., 2025). We use these works as methodological precedents.

Our setup differs from standard single-text classification in two ways. First, cue labeling is relational and thread-aware: the same lexical form can be parasocial or non-parasocial depending on whether it is targeted to OP and whether it seeks relational closure. Second, we explicitly incorporate nullification tests (placebo outcomes, negative controls, prevalence-preserving random labels), which are less common in standard annotation benchmarks but necessary here due to the absence of external ground truth labels for AI-AI PSR.