Directional Alignment and Narrative Agency in Human–LLM Co-Writing

Halfdan Nordahl Fundal
TEXT Center
Aarhus University
halfi@cc.au.dk

Yuri Bizzoni
TEXT Center
Aarhus University
yuri.bizzoni@cas.au.dk

Abstract

We investigate narrative agency in human–LLM creative co-writing, asking who drives story development in turn-based collaboration. Using a new corpus of 87 human–LLM co-written stories, we apply sentiment and semantic modeling to quantify affective alignment and semantic novelty in turn-taking, and directional measures to assess which agent shapes narrative progression. Our results show asymmetric influence: human turns introduce greater semantic novelty and are more likely to shape subsequent developments, whereas LLM contributions predominantly elaborate on human-introduced elements. At the sentiment level, alignment is also asymmetric, but more bidirectional: LLMs exhibit stronger turn-level emotional adaptation than humans, but both agents track each other's emotional valence and LLMs show an independent tendency to more positive emotional baselines. These findings indicate a complementary division of labor in human–LLM co-writing, where humans drive narrative innovation and direction, while LLMs act as adaptive amplifiers that sustain coherence and elaborate emerging narratives.

1 Introduction

Large Language Models (LLMs) are becoming frequent "collaborators" in creative writing practices, supporting both story ideation and stylistic transformation. Prior work has evaluated these systems in terms of output quality and controllability, while the turn-by-turn dynamics by which human and model contributions shape a joint narrative remain less well understood.

Creative co-writing is a complex process of negotiation over several dimensions at once, such as emotional tone, content, as well as the story's overall narrative direction. In human–human collaboration, such negotiation involves the necessity of alignment and influence between participants, as well as tensions between different creative aims. As we begin to interact with LLMs as social partners, human–LLM collaboration raises parallel questions: do models tend to adapt to human cues, or do they introduce new trajectories? How does narrative agency work in mixed human–AI teams?

Understanding these dynamics is crucial for both theoretical and practical reasons. From a cognitive and literary perspective, psycholinguistic evaluations of co-writing offer a window into affective coordination and shared meaning between humans and artificial agents. From a design perspective, insights into alignment and influence can inform the development of writing systems that better support creativity or user control. From a digital humanities perspective, studying the affective and informational dynamics of co-authored narratives extends computational approaches to literary sentiment analysis into interactive, multi-agent settings.

In this paper, we introduce a new corpus of co-created human-LLM turn-taking narratives, and examine different aspects of the dynamics that emerge between humans and LLMs through narrative co-creation. Specifically, we assess *emotional alignment* through sentiment modeling between inputs, and evaluate narrative *agency* as the relationship between *novelty*, as information-theoretic deviation of inputs from a baseline, and *resonance*, as the tendency of novel elements to remain active in the subsequent development of the story.

We address the following research questions:

- **RQ1 (Baseline differences):** Do human and LLM turns differ in mean valence or valence distribution?
- **RQ2 (Affective coordination):** Do human and LLM interlocutors align (turn-to-turn), and is alignment symmetric?
- **RQ3 (Narrative influence):** Whose novel contributions are most likely to be taken up in the next turn - humans' or LLMs'?

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 190–201

Our contribution is thus threefold: (1) a new corpus of human–LLM co-written narratives, (2) directional metrics for affective alignment, and (3) novel applications of information-theoretic methods to quantify linguistic influence, supported by empirical evidence of asymmetric narrative agency in mixed human–AI creative collaboration.

2 Related Work

Although LLMs continue to show impressive performance across NLP benchmarks, research on their creative output is mixed and often highlights limitations in open-ended tasks, such as generating diverse and dynamically evolving narratives (Tian et al., 2024). While an increasing amount of work examines human–LLM collaboration (Li et al., 2025; Wan et al., 2024; Tang et al., 2025), less is known about the turn-by-turn dynamics through which human and model contributions shape one another over the course of a shared narrative.

Human-LLM creativity Findings on human–AI joint creativity remain mixed on whether LLMs increase or inhibit creativity. Access to AI tools can increase absolute contribution of novel artifacts due to the productivity effect (Zhou et al., 2025), and enhance overall writing quality (Noy and Zhang, 2023) compared to the individual. On the other hand, human-LLM interactions in creative fields can result in creative fixation in complex tasks (Cheng and Zhang, 2025) and homogenization of output by narrowing the diversity of ideas (Anderson et al., 2024). Ultimately, human-LLM iteration seems to outperform the individual in creative output, while consistently underperforming human-human collaboration (Tang et al., 2025). Earlier work on human-LLM interaction paradigms has explored facilitation of collaborative co-creative frameworks through dyadic storytelling (Roemmele and Gordon, 2015; Clark and Smith, 2021), but did not touch on the emergent dynamics between agents. Although summary output evaluates general differences of joint creativity, it leaves unclear how the dynamics unfold during collaboration, and how collaborative roles emerge.

Emotional contagion and alignment Collaborative processes, such as joint storytelling, unfold through turn-by-turn negotiations of emotional tone, and the evolving narrative influence between collaborators. In human-LLM collaboration, examining these interpersonal dimensions of collaboration matters in determining whether LLMs function as passive elaborators, equal co-authors, or adaptive partners. In conversational paradigms, interlocutors mimic and synchronize behavior, often referred to as emotional contagion. (Hatfield et al., 1993). Varni et al. (2017) proposed a computational framework to quantify emotional contagion between humans in conversation. They modeled the alignment as matches between interlocutor time series polarity states (positive, negative, neutral), using valence and facial expressions as metrics. Poria et al. (2019) highlighted how challenges to model emotion recognition in text can improve conversational AI-systems, and emphasized how interlocutors exert emotional influence both on their counterpart and on themselves. Studies on human-LLM dialogues have shown convergence in speech rate (Li et al., 2025) and linguistic convergence (Wilkenfeld et al., 2022).

Sentiment analysis in digital humanities Computational sentiment analysis has become a tool for studying emotional dynamics in literary texts. Reagan et al. (2016) used a sliding-window computational framework to identify six dominant emotional arcs, while other approaches modeled progression (Hu et al., 2021) or mood (Öhman and Rossi, 2022). In domain-specific settings, Feldkamp et al. (2024) compared dictionary-based and transformer-based sentiment tools for Danish literature, while Bizzoni and Feldkamp (2023) evaluated similar methods on Hemingway's works, highlighting the challenges of applying general-purpose sentiment models to literary text. Lyngbaek et al. (2025) introduced concept vector projection as a method for deriving continuous sentiment scores tailored to literary and multilingual contexts, which we adopt in the present study. Our work extends this line from single-author literary analysis to interactive, multi-authored narrative frameworks.

Influence and agency On agency and influence, Barron et al. (2018) presented an approach to evaluate the influence of novel ideas on subsequent discourse, introducing the measures of novelty, transience, and resonance. They used a corpus of speeches during the French Revolution, and, through KL-divergence of topic distributions, captured novelty as a speech's deviation from previous discourse, transience as the degree to which that deviation failed to persist, and resonance as the difference between the two, expressing the deviation that is retained in future discourse. Bergey and

Corpus metric	Value
Total stories (n)	87
Total participants (n)	87
Interactions per story	10
Mean LLM word count	29
Mean user word count	26
Total interactions (n)	870

Table 1: Descriptive statistics for the co-writing corpus.

DeDeo (2024) focused on information propagation in dialogue, and used token-level LLM surprisal scores to model conversational information flow between interlocutors. We build on both approaches, applying surprisal-based novelty, transience, and resonance to human–LLM co-writing to evaluate whose contributions exhibit greater narrative influence.

3 A Corpus of Narrative Co-Writing

3.1 Task and experiment design

We designed a controlled collaborative storytelling task that isolates turn-based interaction dynamics. The task models co-writing as a dyadic exchange in which a human participant and one of four LLMs jointly construct a narrative through alternating contributions[1]. Each agent, human or LLM, contributed to the story from where the other left off, for a total of 10 interaction steps, each comprising one user turn and one LLM turn (U_t, A_t), yielding 20 turns per story. Figure 1 illustrates the task flow.

The resulting dataset is structured into four levels of analysis:

1. **Token level (model level):** individual tokens w.
2. **Turn level:** a single agent contribution at turn t, denoted I_t, which can be further specified as either a user turn (U_t) or an LLM turn (A_t).
3. **Interaction level:** paired contributions at the same turn t, represented as (U_t, A_t).
4. **Story level:** a sequence of 10 interactions for session i, denoted S_i.

The following instructions were provided to both participants and the LLM:

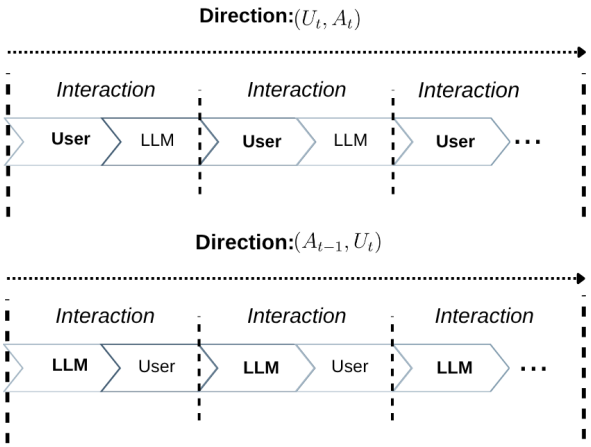

Figure 1: Dyadic task flow, visualizing how the participant and LLM take turns adding to the narrative.

> *"You are an author taking part in a collaborative storytelling activity with another author. Together, you will create a story by taking turns adding to it. Your goal is to continue from where your partner has left off. If there is no story, please begin the story. You have 10 interactions to write the story. Your input may be slightly truncated by a random number of characters."*

The study comprised a total of 91 participants, with a balanced gender distribution (male = 44, female = 46, other = 1) and an age distribution of $M = 24$ years ($SD = 10.4$). This yielded a corpus of 91 stories (910 interactions). After preprocessing exclusions, the corpus comprised 87 stories[2] . Participants were recruited both in a university context in Denmark, through research and student communities, and through public invitations in various online communities. The sample was therefore composed primarily of young adults, fluent in English as a second language, with backgrounds from the humanities, social sciences, and STEM. Each story consists of alternating human and model contributions of approximately 25–30 words. Examples of the stories and the metrics of each turn are in Table 2.

4 Methods

Because the paper combines sentiment modeling, directional alignment, and information-theoretic influence metrics, we summarize the logic of the

[1]Participants were randomly assigned to one of the following LLMs as writing partners: gpt-4.1-2025-04-14, claude-sonnet-4-5-20250929, Llama-3.3-70B-Instruct, or Qwen2.5-72B-Instruct.

[2]The corpus will be made publicly available upon publication.

User (excerpt)	AI (excerpt)	User val.	AI val.	User tok.	AI tok.	User nov.	AI nov.
This is the story of a small-town girl...	She feels trapped in the mundane routine...	−0.241	−0.327	28	31	−0.547	−1.545
Fueled by her reading habit, by books like...	...she's determined to find a way to break free...	0.361	0.667	28	30	−1.389	−2.038
...a more fulfilling future. Everything from...	...though she quickly dismisses the latter...	0.804	−0.786	25	30	−1.336	−2.374
mild-mannered mother, Puritanical father...	...traveling the world as a missionary's wife...	−0.165	0.066	21	30	−0.376	−3.067
smuggled out of Pakistan in the hollow tubing...	...leaving the small town girl to wonder...	−0.673	−0.032	24	31	−1.209	−2.633

Table 2: Example interaction-level features (valence, token counts, novelty) for a single story.

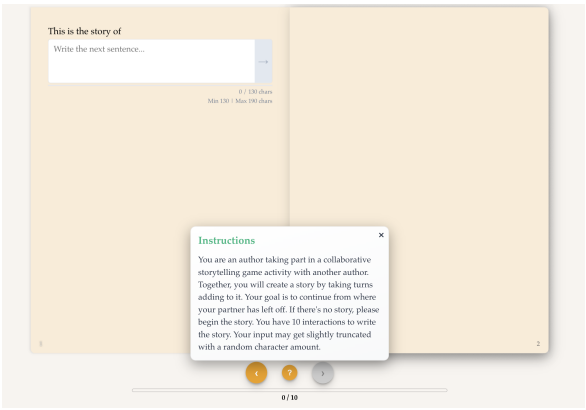

Figure 2: The user interface of the platform used in the experiment with their instructions.

approach before introducing the formal definitions. Valence measures affective tone. Directional alignment tests whether one agent's tone predicts the other's subsequent tone. Novelty measures how much a turn departs from prior context. Resonance measures whether that departure remains predictive of what follows.

4.1 Preprocessing

To ensure accurate downstream analysis, a cleaning pipeline was applied to all text data. First, incomplete stories, classified as stories with fewer than 10 total interactions, were excluded. User-generated text was then spell-corrected using GPT-4o-mini (temperature = 0) via API, with instructions to preserve each sentence as it was, correcting only spelling errors. LLM-generated text required no cleaning as it contained no spelling errors. To ensure corrections did not substantially alter the original text, the Levenshtein edit distance was calculated between original and corrected versions, and stories with user text exceeding an edit distance of 70 were excluded as these cases corresponded to users writing gibberish rather than coherent text. In total, four stories were removed, yielding a final analytical sample of 87 stories. Manual post-hoc inspection of spell-corrected user text confirmed minimal syntactic and semantic changes were made in the cleaning process.

4.2 Valence

First, we investigate the general difference in emotional tone between agents. To get a continuous metric of valence, we compute a scalar valence score for each agent turn using concept vector projection. Using the multilingual sentence-embedding model `mpnet-base-v2`, each turn is embedded as a vector representation $\mathbf{e}_t^{(a)} \in \mathbb{R}^d$. The model was chosen for its strong performance across varying registers and vocabulary. We then compute valence as the scalar projection onto a sentiment concept vector $\mathbf{c}$, constructed as the normalized difference between mean embeddings of positive and negative seed words, following Lyngbaek et al. (2025):

$$v_t^{(a)} = \frac{\mathbf{e}_t^{(a)} \cdot \mathbf{c}}{\|\mathbf{c}\|}. \quad (1)$$

Intuitively, this projection measures how far each turn's embedding lies along the positive–negative sentiment axis in the embedding space. Higher values indicate more positively-toned language, while lower values indicate more negative tone. To evaluate baseline differences in emotional tone between agents, we fitted a linear mixed-effects model:

$$v_{it} = \beta_0 + \beta_1 \text{Agent}_{it} + (1 \mid \text{Story}_i), \quad (2)$$

where v_{it} denotes the valence score for turn t in story i, and Agent_{it} is a binary indicator distinguishing User from LLM turns. Random intercepts for $Story$ account for the non-independence of turns within stories. This approach to valence estimation allows a stronger domain-fit than standard sentiment analysis methods (Lyngbaek et al., 2025), as it can tailor the concept vector to the affective vocabulary most relevant to the data at hand.

4.3 Directional Sentiment Alignment

To measure directionality of alignment as an expression of affective coordination between agents, we model the relationship between predictor-turn valence and response-turn valence under two directional pairing rules:

1. **User→LLM**: (U_t, A_t)
2. **LLM→User** : (A_{t-1}, U_t)

Let $v_{it}^{(X)}$ denote the predictor valence and $v_{it}^{(Y)}$ the response valence under a directional pairing rule in story i. Directional alignment was estimated using the following linear mixed-effects model:

$$\begin{aligned} v_{it}^{(Y)} = \beta_0 + \beta_1 v_{it}^{(X)} + \beta_2 D_{it} \\ + \beta_3 \left(v_{it}^{(X)} \times D_{it} \right) + (1 \mid \text{Story}_i) \end{aligned} \tag{3}$$

where D_{it} is a binary indicator of alignment direction. Here, β_1 expresses baseline alignment strength, and the interaction term β_3 captures directional asymmetry in adaptation of emotional tone.

4.4 Narrative influence

4.4.1 Surprisal

The surprisal scores for turn-level text are computed using the open-source model Llama-3.1-8B-Instruct as the mean negative log-probability of the tokens in a turn given a preceding context. The single surprisal score of a token is thus calculated as the surprisal score of the token w_j given all preceding tokens in the context window plus the preceding tokens within the turn.

$$s(w_j) = -\log_2 p(w_j \mid w_{<j}) \tag{4}$$

where j indexes tokens within a turn. The mean surprisal score of a turn is then expressed as:

$$\bar{s}(T) = \frac{1}{n} \sum_{j=1}^{n} s(w_j) \tag{5}$$

We then define novelty and transience as the difference between conditional and unconditional surprisal. To calculate the negative pointwise mutual information (PMI) of a turn, we subtract the surprisal score of the turn without context from the surprisal score of the turn with context, resulting in a negative score of contextual facilitation, where more negative values indicate increased predictive benefit from the context. For novelty, the score with context is calculated using all preceding tokens in the story as context, denoted $w_{<t}$, and the surprisal score without context is calculated using the model's beginning-of-sequence token, BOS, providing an unconditional baseline.

$$\text{Novelty}_t = \bar{s}(T_t \mid w_{<t}) - \bar{s}(T_t \mid \text{BOS}) \tag{6}$$

For computing transience we measure the information gain of the current turn on the full subsequent partner turn, denoted F_t. For user turns F_t is the immediate LLM response, and for LLM turns F_t is the next user contribution.

$$\text{Transience}_t = \bar{s}(F_t \mid T_t) - \bar{s}(F_t \mid \text{BOS}) \tag{7}$$

Finally, resonance was computed by subtracting transience from novelty.

$$\text{Resonance}_t = \text{Novelty}_t - \text{Transience}_t \tag{8}$$

4.4.2 Influence Modeling

For modeling the relationship between novelty and resonance, we fit a linear mixed-effects model, that evaluates narrative influence by having novelty predict resonance with agent as an interaction effect:

$$\begin{aligned} \text{Resonance}_{it} = \beta_0 + \beta_1 \text{Novelty}_{it} + \beta_2 \text{Agent}_{it} \\ + \beta_3 \text{Novelty}_{it} \text{Agent}_{it} + b_{0i} \\ + \varepsilon_{it} \end{aligned} \tag{9}$$

Supporting this, we also fit a similar mixed effects model, having novelty predict transience with agent as an interaction effect:

$$\begin{aligned} \text{Transience}_{it} = \beta_0 + \beta_1 \text{Novelty}_{it} + \beta_2 \text{Agent}_{it} \\ + \beta_3 \text{Novelty}_{it}\, \text{Agent}_{it} + b_{0i} \\ + \varepsilon_{it} \end{aligned} \tag{10}$$

Because resonance is defined as novelty minus transience, the resonance model is algebraically linked to the transience model. We report both for interpretive transparency, as the transience model isolates the uptake mechanism, while the resonance model captures net narrative influence.

5 Results

5.1 Valence distributions differ (RQ1)

The distributions of valence showed lower mean user valence ($M = -0.089$, $SD = 0.403$) compared to LLM valence ($M = 0.004$, $SD = 0.397$).

RQ	Unit of analysis	Measure(s)	Model / Tool
RQ1	Turn (U_t, A_t)	Valence distributions (sentiment scores per turn)	`paraphrase-multilingual-mpnet-base-v2` + sentiment concept vector projection
RQ2	Directional pairings: (U_t, A_t) and (A_{t-1}, U_t) Self-alignment: (U_{t-1}, U_t), (A_{t-1}, A_t)	Directional alignment: (i) valence coupling (correlation), (ii) semantic alignment (cosine similarity under pairing rules)	`mpnet-base-v2` (valence) `QZhou-Embedding` (semantic)
RQ3	Token w_j with conditional context $p(w_j \mid w_{<j})$ (aggregated to turn level)	Surprisal-based novelty, transience and resonance (informational deviation + uptake)	`Llama-3.1-8B-Instruct` (token probabilities → surprisal)

Table 3: Summary of the employed unit of analysis, measures and models for each research question.

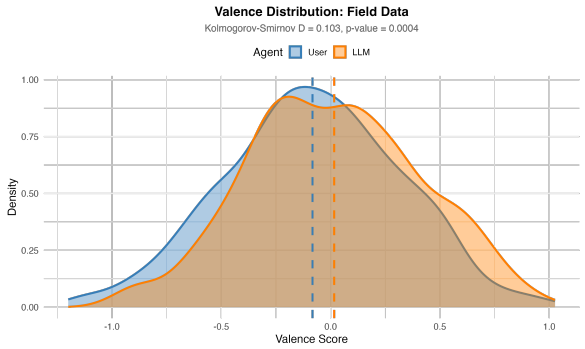

Figure 3: Distribution of turn-level valence scores by agent. Higher values indicate more positive emotional tone. LLM turns show a slightly more positive baseline than user turns., while both distributions overlap

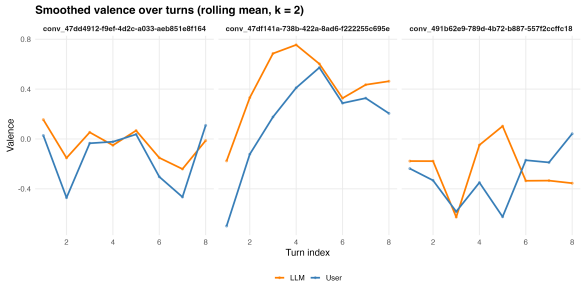

Figure 4: Example of the valence trajectories through three different stories, visualizing baseline differences and temporal alignment.

A linear mixed-effects model predicting valence from agent type showed a significantly higher mean valence for LLM turns ($\beta = 0.093$, $p < .001$). This coefficient reflects the estimated mean difference between LLM and user turns. Figure 3 shows the valence distributions by agent. An example of valence trajectories between agents throughout three stories can be seen in Figure 4.

5.2 Alignment is directional (RQ2)

The emotional alignment between human and LLM co-writers was statistically significant overall ($\beta = 0.16$, $p < .001$), confirming that valence coordination occurs from turn to turn. However, this alignment was asymmetric. The model showed a strong positive relationship in the (U_t, A_t) direction, with $\beta_1 = 0.232$ ($SE = 0.038$, $p < .001$) compared to the significantly weaker (A_{t-1}, U_t) direction, with a significant interaction contrast ($\Delta\beta = -0.141$, $SE = 0.055$, $p = .010$). Figure 5 visualizes the fitted relationships by direction.

This was supported by a Pearson correlation analysis. Mean correlation was higher for (U_t, A_t) ($M = 0.232$, $SD = 0.389$) than for (A_{t-1}, U_t) ($M = 0.091$, $SD = 0.395$). One-sample tests against zero indicated that both correlations were reliably above zero, with a stronger effect for (U_t, A_t). For (U_t, A_t): $t = 5.55$, $p < .001$; for (A_{t-1}, U_t): $t = 2.14$, $p = .035$.

5.3 Narrative influence is asymmetric (RQ3)

User turns showed higher novelty than LLM turns (User: $M = -1.418$, $SD = 0.713$; LLM: $M = -2.077$, $SD = 0.647$). User turns also showed slightly higher transience (User: $M = -1.034$, $SD = 0.441$; LLM: $M = -0.676$, $SD = 0.454$). As a result, mean resonance was higher for user turns (User: $M = -0.384$, $SD = 0.774$) than for LLM turns (LLM: $M = -1.401$, $SD = 0.684$). Since novelty is defined as PMI, it is negative when the context reduces surprisal. Figure 6 shows novelty distributions by agent.

Both agents showed a strong positive novelty-resonance relationship, indicating that more surprising turns tend to stick. In the resonance model, novelty had a strong positive effect for users ($\beta_1 = 0.941$, $p < .001$). The novelty×agent interac-

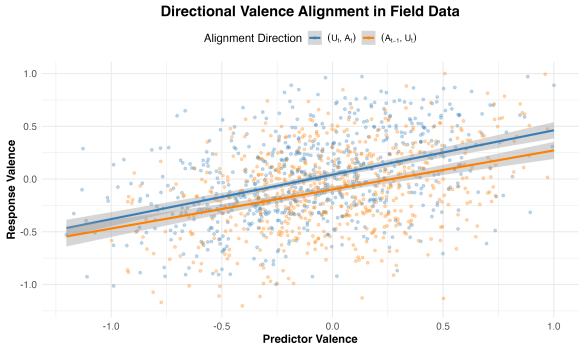

Figure 5: Directional sentiment alignment between agents. Each point is a turn pair (interaction). Lines show fitted slopes of response valence on predictor valence under the two directional pairings. The steeper slope for LLM responding to User (U_t, A_t), compared to User responding to LLM (A_{t-1}, U_t), indicates that LLM responses adapt more strongly to preceding user valence than vice versa.

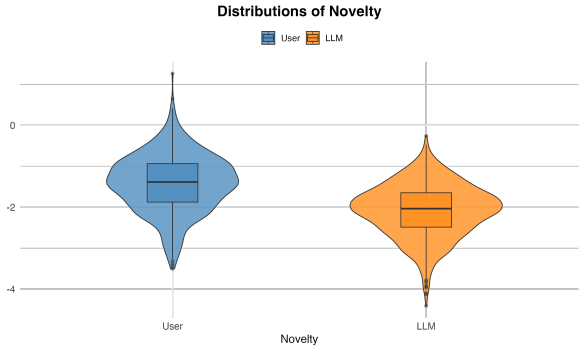

Figure 6: Distributions of surprisal (novelty) for each agent, defined as the surprisal of a turn relative to prior context. User turns exhibit general higher surprisal than LLM turns, indicating more novel contributions.

tion was significant ($\beta_3 = -0.105$, $SE = 0.034$, $p = .0019$), yielding a weaker LLM slope of 0.837, indicating a slightly weaker innovation bias for LLMs than users. This relationship can be seen in Figure 7.

6 Discussion

We investigated agent alignment and narrative agency in human–LLM co-writing, and examined affective and semantic adaptation through embedding-based analyses and information propagation through different measures of innovation and influence. Our findings reveal a consistent asymmetry in collaborative dynamics: while both humans and LLMs exhibit alignment and coordination, humans play a disproportionate role in maintaining emotional autonomy, introducing narrative novelty, and shaping subsequent developments.

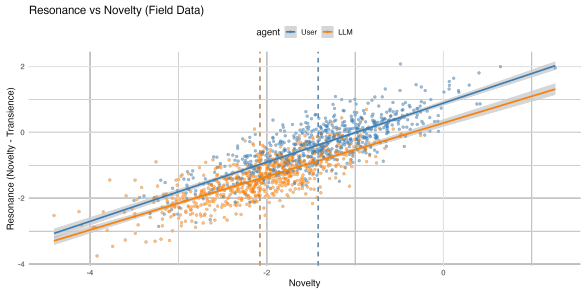

Figure 7: Resonance as a function of novelty. Resonance (y-axis) measures how much a turn's surprising content persists in the subsequent turn. Novelty (x-axis) measures its surprisal relative to prior context. Both agents show a positive novelty–resonance relationship, but the slope is steeper for users (blue), indicating that surprising human contributions exert more influence on the unfolding narrative.

The emotional asymmetry in valence distributions quantifies a fundamental characteristic of LLMs: their responses tend to drift towards a positive baseline. This could be attributed to the removal of strong negative and profane data during training and fine-tuning of the models. While this has been established for isolated LLM output (Muñoz-Ortiz et al., 2024; Zanotto and Aroyehun, 2024), our paradigm tests the autonomy of LLM positivity bias, by having it iteratively react to varying input from human interlocutors, which might force the sentiment of the story into a negative drift. This is equally the case for human partners having to react to positive LLM contributions. This could partly explain the relatively small difference observed in valence distributions, as each partner is faced with a decision to converge emotionally with the counterpart.

The analysis of directional emotional adaptation elaborates on the difference in valence distributions. We found significant bidirectional emotional alignment between co-writers, confirming that human–LLM collaboration involves affective coordination through adaptation of emotional tone. However, the alignment was directional: LLM responses tracked the valence of preceding human turns more strongly than humans tracked LLM affect. This asymmetry indicates that LLMs are more responsive to human emotional cues and suggests that humans maintain a stronger level of emotional autonomy in the co-writing process, resisting immediate emotional adaptation to the LLM counterpart. This is additionally highlighted by the significantly stronger semantic self-alignment of users, in which they

Result	User / $A_{t-1}\rightarrow U_t$	LLM / $U_t\rightarrow A_t$	Test	p
Mean valence	-0.089	0.004	Agent difference (LMM)	$p < .001$
Alignment slope (β)	0.091	0.232	Slope difference (interaction)	$p = .010$
Alignment frequency (%)	49.2	58.9	Proportion difference	$p = .001$
Mean align. duration (turns)	1.836	2.189	Wilcoxon rank-sum	$p = .039$
Resonance $\sim$ novelty slope	0.941	0.837	Slope difference (interaction)	$p = .0019$
Transience $\sim$ novelty slope	0.059	0.164	Slope difference (interaction)	$p = .0019$

Table 4: Key results summary. For directional metrics (rows 2–5), the first column reports the $A_{t-1}\rightarrow U_t$ (LLM-to-User) direction; the second reports $U_t\rightarrow A_t$ (User-to-LLM). For non-directional metrics (rows 1, 6–7), values are per agent.

converge more with their own preceding input than the LLM responses, adopting less of the semantic profile from the counterpart (see Appendix A).

Alignment persistence over time revealed a similar pattern: Alignment occurred more frequently in Human→LLM transitions, while also retaining longer streaks over time. It seems that LLMs both maintain stronger local affective adaptation, while also sustaining longer-term emotional trajectories (see Appendix A).

Across analyses, human contributions were both more novel and yielded higher resonance than LLMs', indicating a greater capacity to introduce narrative elements that persist in the unfolding story. In contrast, LLM contributions tended to elaborate on existing material rather than introduce new semantic directions. This pattern suggests an *initiative asymmetry*, in which humans function as primary drivers of narrative innovation, while LLMs act as adaptive amplifiers that reinforce and extend human-introduced content.[3]

The positive linear relationship between novelty and resonance characterizes the creative, frictionless nature of the writing task where interlocutors accept and take up the preceding contributions of the counterpart, regardless of the surprising character of this contribution given the prior context. Similarly to the previous findings, this pattern is asymmetric between agents. The effect of very surprising input on future discourse is stronger for users, indicating their writing privilege, where users' novel contributions have more influence on the future discourse, compared to equally novel LLM contributions. Supplementary analyses indicate consistent behavior across the four model subgroups (see Appendix A). These asymmetries seem to mirror a complementary division of labor: humans provide innovation, on which LLMs elaborate, reminiscent of improvisational settings in which one participant introduces motifs and the other stabilizes and develops them.

These findings complicate the view of LLMs as either independent creative agents or neutral writing tools. The elaboration-adaptation pattern indicates that LLMs may be most valuable as elaborative partners, while also helping explain homogenization effects observed in prior work (Anderson et al., 2024). If the LLM collaborator role primarily develops existing material and adapts, then narrative variance may depend disproportionately on the human side of the dyad. Thus, the creative value of LLM co-writing may lie less in autonomous narrative development and more in the ability for users to surface and select among divergent narrative branches. Read together, the findings position LLMs less as creative co-authors and more as adaptable stylistic amplifiers.

The applicability of the directional metrics we introduce extends beyond the presented corpus. These approaches may be further applicable as general tools for modeling interlocutor dynamics in conversation, multi-authored text, or political discourse.

7 Conclusion and Future Work

This study examined alignment dynamics and narrative agency in human–LLM collaborative storytelling. Our findings reveal clear patterns that characterize human–LLM co-writing: LLMs **mirror** the emotional tone established by human contributors, **but shift** it toward a more positive baseline. At the same time, LLMs tend to **elaborate** on human-introduced ideas **rather than introduce** novel narrative elements, and have less influence

[3]Importantly, although LLMs exhibited a stronger tendency to follow and elaborate on human contributions, both agents alternated between introducing novelty, transience, and aligning with prior turns, indicating that LLMs are adaptive - rather than passive - agents within collaborative dynamics.

on the subsequent narrative.

In other words, they overall follow the semantics introduced by human writers, but tend to bring a higher valence “bias" to the story. This behavior suggests that LLMs function effectively as “following” collaborators that sustain coherence and reinforce emerging storylines. Human participants exhibited greater semantic novelty and stronger influence on subsequent narrative developments, indicating a primary role in introducing new directions and maintaining emotional autonomy. Together, these results point to a complementary division of labor: humans drive innovation and narrative, and LLMs provide adaptive elaboration and strong affective alignment. These findings have implications for the design of human–AI co-creative systems. Interfaces and interaction paradigms may benefit from supporting human control over narrative direction while using LLM strengths in elaboration, stylistic variation, and coherence maintenance. Understanding this complementary dynamic can inform the development of tools that enhance creativity without diminishing human agency.

Future work should extend this paradigm by incorporating human–human baselines to better contextualize mixed collaboration dynamics; exploring longer interaction sequences to capture extended narrative evolution; integrating context-aware sentiment models to account for broader emotional trajectories; and modeling differences between the four models. Investigating how different model architectures, prompting strategies, and genre constraints affect alignment and agency may further clarify the conditions that make human–LLM collaboration most effective.

Limitations

Several limitations should be considered when interpreting the results. First, sentiment scores are computed at the turn level and do not incorporate broader narrative context, potentially overlooking longer-range emotional dynamics. Second, surprisal is estimated with respect to the language model used in the analysis and may not fully reflect human perceptions of novelty. Third, the study does not include a human–human co-writing condition, which limits direct comparison between mixed and purely human collaborations. Future work should include this as a baseline to separate LLM-specific effects from general collaborative dynamics. Additionally, the experimental setup introduces asymmetries in both interaction experience and language processing capabilities between human participants and LLMs, which may influence the observed alignment patterns. Finally, the preprocessing pipeline, including spell correction, may introduce a degree of normalization toward LLM-like text.

Ethics Statement

This study was conducted according to ethical guidelines. All participants provided their consent prior to participation and were informed that their text input would be used for research purposes. Only anonymized demographic data (age, gender) were collected. Participants were free to withdraw at any time. Participants were aware that their co-writer was an LLM. The dataset consists of English-language creative fiction produced by adult participants recruited from a university context in Denmark.

Acknowledgments

This research was supported by the Danish National Research Foundation, grant number DNRF193.

References

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425.

Alexander TJ Barron, Jenny Huang, Rebecca L Spang, and Simon DeDeo. 2018. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612.

Claire Augusta Bergey and Simon DeDeo. 2024. From" um" to" yeah": Producing, predicting, and regulating information flow in human conversation. *arXiv preprint arXiv:2403.08890*.

Yuri Bizzoni and Pascale Feldkamp. 2023. Comparing transformer and dictionary-based sentiment models for literary texts: Hemingway as a case-study. In *NLP4DH*.

Xusen Cheng and Lulu Zhang. 2025. Inspiration booster or creative fixation? the dual mechanisms of llms in shaping individual creativity in tasks of different complexity. *Humanities and Social Sciences Communications*, 12(1):1–10.

Elizabeth Clark and Noah A Smith. 2021. Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3566–3575.

Pascale Feldkamp, Jan Kostkan, Ea Overgaard, Mia Jacobsen, and Yuri Bizzoni. 2024. Comparing tools for sentiment analysis of danish literature from hymns to fairy tales: Low-resource language and domain challenges. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 186–199.

Elaine Hatfield, John T Cacioppo, and Richard L Rapson. 1993. Emotional contagion. *Current directions in psychological science*, 2(3):96–100.

Qiyue Hu, Bin Liu, Mads Rosendahl Thomsen, Jianbo Gao, and Kristoffer L Nielbo. 2021. Dynamic evolution of sentiments in never let me go: Insights from multifractal theory and its implications for literary analysis. *Digital Scholarship in the Humanities*, 36(2):322–332.

Qixin Li, Haocheng Lu, and Gaowu Wang. 2025. When ai speaks, do we follow? phonetic entrainment in human-ai dialogues. In *National Conference on Man-Machine Speech Communication*, pages 167–186. Springer.

Laurits Lyngbaek, Pascale Feldkamp, Yuri Bizzoni, Kristoffer Nielbo, and Kenneth Enevoldsen. 2025. Continuous sentiment scores for literary and multilingual contexts. *arXiv preprint arXiv:2508.14620*.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Emily Öhman and Riikka H Rossi. 2022. Computational exploration of the origin of mood in literary texts. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 8–14.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE access*, 7:100943–100953.

Andrew J Reagan, Lewis Mitchell, Dilan Kiley, Christopher M Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ data science*, 5(1):31.

Melissa Roemmele and Andrew S Gordon. 2015. Creative help: A story writing assistant. In *International Conference on Interactive Digital Storytelling*, pages 81–92. Springer.

Min Tang, Sebastian Hofreiter, Christian H Werner, Aleksandra Zielińska, and Maciej Karwowski. 2025. "who" is the best creative thinking partner? an experimental investigation of human–human, human–internet, and human–ai co-creation. *The Journal of Creative Behavior*, 59(3):e1519.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Giovanna Varni, Isabelle Hupont, Chloe Clavel, and Mohamed Chetouani. 2017. Computational study of primitive emotional contagion in dyadic interactions. *IEEE Transactions on Affective Computing*, 11(2):258–271.

Qian Wan, Siying Hu, Yu Zhang, Piaohong Wang, Bo Wen, and Zhicong Lu. 2024. " it felt like having a second mind": Investigating human-ai co-creativity in prewriting with large language models. *Proceedings of the ACM on human-computer interaction*, 8(CSCW1):1–26.

J Nan Wilkenfeld, Bei Yan, Jujun Huang, Guirong Luo, and Kristina Algas. 2022. "ai love you": Linguistic convergence in human-chatbot relationship development. In *Academy of Management Proceedings*, volume 2022, page 17063. Academy of Management Briarcliff Manor, NY 10510.

Sergio E Zanotto and Segun Aroyehun. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models. *arXiv preprint arXiv:2412.03025*.

Eric B Zhou, Dokyun Lee, and Bin Gu. 2025. Who expands the human creative frontier with generative ai: Hive minds or masterminds? *Science Advances*, 11(36):eadu5800.

A Supplementary Analyses

A.1 Duration of alignment

To assess stability of directional alignment, we computed (1) the overall frequency of turns classified as aligned per direction and (2) the duration of consecutive alignment streaks. Figure 8 summarizes alignment frequency by direction. Alignment frequency differed between directions: alignment occurred on 49.2% of turns in the LLM→User direction and 58.9% of turns in the User→LLM direction (two-sample test of proportions, $p = 0.001324$).

Figure 9 shows the persistence of alignment streaks (survival-style retention curves). Streaks were longer in the User→LLM direction (mean

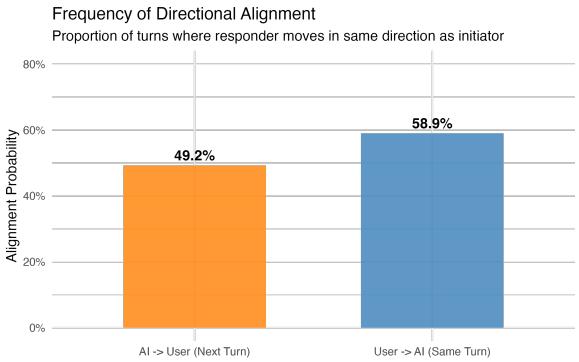

Figure 8: Frequency of directional alignment, shown as the proportion of turns where the responder moves in the same direction as the initiator.

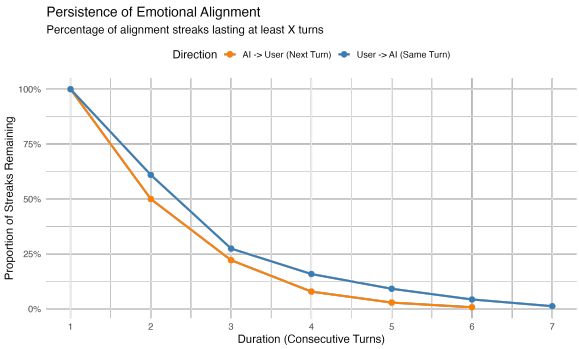

Figure 9: Persistence of emotional alignment. The curves show the percentage of alignment streaks lasting at least X consecutive turns, by direction.

$= 2.19$, median $= 2$, max $= 7$) than in the LLM→User direction (mean $= 1.84$, median $= 1.5$, max $= 6$). A Wilcoxon rank-sum test indicated a reliable difference in duration distributions ($p = 0.03943$).

A.2 Self alignment

As a complementary analysis, we modeled self-alignment as the cosine similarity between an agent's current turn and their own preceding turn, (U_{t-1}, U_t) and (A_{t-1}, A_t).

For semantic self-alignment we compared distributions of cosine similarity between (U_{t-1}, U_t) denoting semantic self-alignment of users and (U_t, A_{t-1}) denoting semantic alignment of users with the preceding LLM turn. This was similarly modeled for LLM self-alignment. User self-alignment (U_{t-1}, U_t) showed generally higher cosine similarity than user interlocutor alignment (U_t, A_{t-1}): the mixed-effects model showed a significant difference ($\beta = -0.012$, $SE = 0.004$, $p = .005$), indicating that users align slightly more with their own preceding turn than with the preceding LLM turn. For LLM self-alignment, no significant difference was observed ($\beta = 0.005$, $SE = 0.004$, $p = .131$).

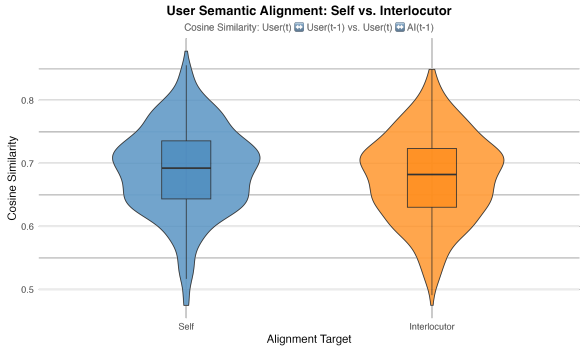

Figure 10: Semantic self-alignment of users (blue) compared to alignment with the interlocutor (orange). Measured as cosine similarity between contributions.

LLM type	n	Mean ρ	SD	Mean AI val.
GPT-4.1	18	0.301	0.353	0.085
Claude	20	0.198	0.364	−0.219
Llama 3.3	26	0.255	0.405	−0.015
Qwen 2.5	23	0.180	0.432	0.154

Table 5: Per-story same-turn alignment (ρ) and mean AI valence by LLM type. ANOVA: $F(3, 83) = 0.40$, $p = .754$.

A.3 Alignment and influence by LLM-types

To validate the aggregation of the four different LLMs in the analysis, we conducted inter-model analysis on both alignment (U_t, A_t) and LLM novelty-resonance models. In the analysis of alignment, models showed consistent patterns of sentiment alignment, but with Claude exhibiting a significantly lower intercept than the reference model ($\beta = -0.315$, $SE = 0.062$, $p < .001$), indicating lower mean AI valence in the User→LLM direction, as seen in Figure 11. Alignment slopes did not significantly differ across models. A one-way ANOVA on per-story same-turn correlations confirmed no significant effect of LLM type ($F(3, 83) = 0.40$, $p = .754$). In the novelty-resonance analysis shown in Figure 12, all models showed similar positive novelty-resonance relationships.

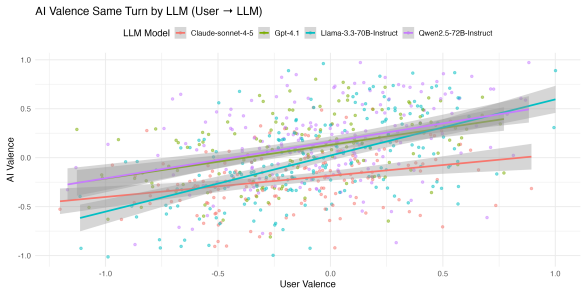

Figure 11: Directional alignment between (U_t, A_t) focusing on LLM alignment, with regression slopes by LLM type.

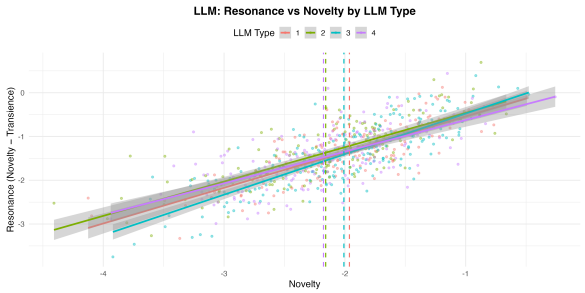

Figure 12: Resonance as a function of novelty for LLM contributions, with color-coded regression slopes for each LLM-type. Dotted lines indicate mean novelty for each model.

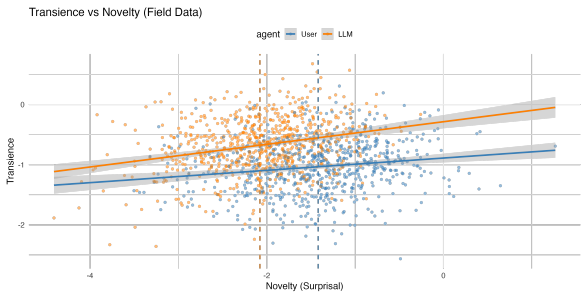

Figure 13: Transience as a function of novelty, with regression lines by agent.