Scholars Journal of Arts, Humanities and Social Sciences

OPEN ACCESS

Abbreviated Key Title: Sch J Arts Humanit Soc Sci
ISSN 2347-9493 (Print) | ISSN 2347-5374 (Online)
Journal homepage: https://saspublishers.com

How Humans Read Stories in the Age of AI: A Cross-Linguistic Psycholinguistic Study of Narrative Prediction, Emotion, and Voice in Human- Vs. AI-Mediated Literature

Inzimam Ul Haq[1*], Ayaan Ahmad Khan[1], Faisal Khan[1], Sidra Kauser[2]

[1]Department of English, National University of Modern Languages (NUML), Islamabad, Pakistan
[2]Department of English, University of Sargodha, Punjab 54470, Pakistan

DOI: https://doi.org/10.36347/sjahss.2026.v14i02.001 | **Received:** 03.11.2025 | **Accepted:** 08.01.2026 | **Published:** 03.02.2026

***Corresponding author:** Inzimam Ul Haq
Department of English, National University of Modern Languages (NUML), Islamabad, Pakistan

Abstract **Original Research Article**

As AI-mediated writing becomes increasingly visible in literary and digital reading contexts, it is critical to understand how readers experience AI-involved narratives across languages. This study examined whether narrative prediction, emotional engagement, and perceived narrative voice differ between human-authored and AI-mediated short stories in two language cohorts: English and Mandarin Chinese. We constructed 12 tightly matched story pairs (6 per language), controlling for length, sentence count, readability, and baseline lexical properties. A large online sample was recruited (N = 652), with exclusions applied using pre-registered criteria, yielding a final analytic sample of N = 528 (English n = 264; Mandarin n = 264). Narrative prediction was assessed using a Cloze Probability Task. Across languages, AI-mediated texts showed lower cloze predictability than human-authored texts, with a significant main effect of Text Type ($\beta = -0.076$, $p < .001$) and a significant Text Type × Language interaction ($\beta = -0.035$, $p = .009$), reflecting a larger predictability penalty in English. Subjective outcomes showed robust main effects of Text Type for Narrative Engagement ($\beta = -0.414$, $p < .001$) and Emotional Intensity ($\beta = -0.375$, $p < .001$) without cross-linguistic interaction, indicating a consistent experiential reduction across cohorts. Narrative voice exhibited the strongest AI-related penalties across Authenticity, Stylistic Naturalness, and Perspectival Coherence (all $p < .001$), with a language-sensitive interaction for coherence ($\beta = -0.105$, $p = .018$). Moderation analyses revealed that AI familiarity attenuated subjective penalties for engagement, emotion, and voice authenticity/naturalness, but did not significantly moderate cloze predictability. An integrative effect-size synthesis and the Narrative Triad Divergence Index further demonstrated a larger overall AI-related divergence in English (NTDI = 1.03) than Mandarin (NTDI = 0.79). Collectively, these findings suggest a multidimensional "AI reading signature" characterized by robust cross-linguistic reductions in emotional engagement and voice authenticity, alongside language-sensitive disruptions in narrative predictability and perspectival coherence.

Keywords: AI-mediated literature; narrative prediction; cloze probability; narrative engagement; emotional intensity; narrative voice; cross-linguistic psycholinguistics; English and Mandarin reading.

1. INTRODUCTION

1.1 Reading narratives in an AI-saturated world

Reading a story has never been a neutral act. Psycholinguistic research shows that language comprehension is fundamentally predictive: readers continuously anticipate upcoming words, events, and discourse moves, and more predictable inputs are processed faster and with less cognitive effort [1–4]. In parallel, literary and media psychology emphasize that narrative reading is deeply affective and immersive. Green and Brock's narrative transportation framework describes how readers become "lost in a story world," with focused attention, emotional engagement, and vivid mental imagery that can reshape beliefs and memories [5,6].

Against this cognitive–affective background, the last few years have seen an unprecedented change: large language models and other generative systems now routinely draft, rewrite, and co-author narrative texts. Recent estimates suggest that more than half of newly published English-language web articles are AI-generated, at least at the level of surface drafting [19], and surveys of novelists in the UK report that a majority now see AI as a potential replacement threat to human-authored fiction [20]. Public debate has focused intensely on copyright, labour, and originality, while popular essays and empirical work alike warn that heavy reliance on generative models can homogenise language and thought, narrowing stylistic diversity and attenuating cognitive engagement [21].

Taken together, these developments suggest that contemporary readers increasingly encounter stories in which AI has acted as author, co-author, or invisible editor. Yet we still know very little about how such AI mediation interacts with the core psycholinguistic and literary dimensions of reading: prediction, emotion, and voice.

1.2 Prediction, emotion, and voice as coupled dimensions of narrative reading

The predictive turn in psycholinguistics has reframed reading as a process in which the brain continuously generates probabilistic expectations at multiple representational levels from phonology and lexicon to syntax, semantics, and situation models [1,3]. More predictable continuations yield faster reading times, smaller N400 amplitudes, and smoother integration [2,4]. Narratives, with their rich event structures and character arcs, are particularly fertile contexts for such predictions: readers anticipate not only words but also plot turns, emotional shifts, and character decisions.

At the same time, narrative research has shown that emotional engagement and empathy are not mere by-products but central mechanisms of narrative impact. Narrative transportation theory posits that when readers are deeply absorbed, they form strong emotional bonds with characters and may adopt attitudes aligned with the story [5,6]. Meta-analytic work indicates that transported readers show stronger emotional responses and are more susceptible to persuasion, especially when they empathize with protagonists and experience vivid mental imagery [11].

A third, often under-operationalised dimension is narrative voice. Classical narratology, from Genette's analysis of voice, time of narration, and perspective to more recent "toolbox" approaches, treats voice as the configuration of "who speaks?", "from where?", and "through whom do we perceive the story world?" [7,8]. Readers do not encounter text as neutral information; they hear a "textual voice" with their "mind's ear" and see events with their "mind's eye," constructing an implicit social agent behind the words [8]. Perceived voice authenticity, coherence, and stance are therefore likely to shape both predictive expectations and emotional trust.

Conceptually, prediction, emotion, and voice are tightly coupled. A stable, credible narrative voice can guide predictions about how the story will unfold and which emotional cues are relevant. In turn, successful predictions can deepen transportation and empathy, while prediction errors at key moments may produce surprise, suspense, or aesthetic pleasure. A narrative that "sounds" emotionally flat or mechanically patterned may still be linguistically fluent but may fail to sustain the same predictive and affective dynamics. This suggests that AI-mediated changes in style and discourse structure could have downstream consequences for both narrative predictability and emotional resonance.

1.3 AI-mediated narratives and emerging reader-side evidence

Empirical research comparing AI-generated and human-written texts is beginning to appear, but it is still fragmented and mostly focused on surface similarity and detection, rather than rich reader experience. Corpus-based work by Sardinha (2024), for example, shows that GPT-generated texts differ significantly from human texts along Biber's multidimensional register dimensions, with AI outputs often failing to match the distributional patterns of genuine spoken and written registers [9]. In education, comparative studies of AI-generated versus human-written articles and assessment passages report that AI texts can match or exceed human texts on readability and correctness, but may differ in coherence, engagement, and stylistic range [10,11].

At the level of reader perception, several studies suggest that people treat texts differently once AI authorship is suspected or disclosed. Work on AI disclosure in communication indicates that revealing a text as AI-generated can reduce perceived authenticity, trust, and empathy, even when the linguistic quality is kept constant [13]. Medical education research similarly finds that readers can often identify AI-generated explanations as less human-like or less empathetic, and that authorship cues shape their evaluative judgments [12]. A recent essay on empirical reader-response research argues that the distinction between human authorship and AI-generated writing is becoming an important axis for studying how readers ascribe intentionality, emotion, and responsibility to texts [14].

Social and cultural discourse mirrors these concerns. Creatives and academics often report that AI-generated narratives feel emotionally thin or socially hollow, even when grammatically flawless, and some explicitly reject "robotic" storytelling as incompatible with the kind of human connection they seek from narrative art [22]. At the same time, large-scale analyses of online content suggest that AI-produced text tends toward stylistic homogenization and repetition, raising questions about how such patterns might affect readers' expectations and long-term narrative diets [19,21].

However, most of this work:

1. focuses on recognition and evaluation (can readers tell AI from human; do they like it),
2. rarely measures online prediction or fine-grained emotional responses, and
3. almost never considers cross-linguistic variation in reader experience.

There is therefore a clear need for controlled psycholinguistic studies that directly examine how AI mediation shapes prediction, emotion, and voice during narrative reading.

1.4 Cross-linguistic perspectives on narrative processing

Reading is not only a cognitive activity; it is also deeply shaped by language-specific and cultural conventions. Research on cross-linguistic reading and literacy shows that differences in writing systems, morphology, and discourse structure can influence how readers allocate attention, build coherence, and use predictive cues [15–18]. For example, cross-linguistic work on sentence processing and morphological awareness indicates that speakers of different languages recruit partially distinct strategies when anticipating upcoming words or integrating morphologically complex items in context [15,16].

In second-language and bilingual reading, studies highlight that discourse-level operations such as cohesion tracking, inference generation, and perspective-taking can vary with readers' language dominance, proficiency, and prior literacy experience [17,18]. These findings suggest that what counts as a "natural" narrative progression, a "well-formed" voice, or a "plausible" emotional trajectory is not universal, but modulated by linguistic and cultural background.

This has direct implications for AI-mediated literature. Large language models are trained on corpora with uneven language coverage, often dominated by English and particular genres. Their narrative priors may thus implicitly encode Anglophone discourse norms, which could align well with readers in some languages but clash subtly with expectations in others. Cross-linguistic research on narrative processing can therefore illuminate whether AI-mediated texts:

- support comparable prediction dynamics across languages,
- sustain similar levels of emotional engagement, and
- instantiate narrative voices that feel equally authentic and coherent to different readerships.

Yet, to our knowledge, no existing study systematically compares human vs AI-mediated narratives across languages using integrated measures of prediction, emotion, and voice.

1.5 The present study

The present study addresses these gaps by bringing together insights from predictive psycholinguistics, narrative transportation and empathy research, and narratology to examine how humans read stories in the age of AI. We focus on three interrelated dimensions:

- Narrative prediction – operationalised through cloze probabilities and/or online reading measures (e.g., self-paced reading), which index how easily readers anticipate and integrate upcoming narrative content.
- Emotion and narrative empathy – assessed via validated scales of transportation, emotional intensity, and character-related empathy, capturing affective engagement with the story.
- Narrative voice – measured through reader ratings of perceived authenticity, narratorial presence, perspectival coherence, and stylistic naturalness, grounded in narratological theory [7,8].

We compare these dimensions across human-authored and AI-mediated versions of short narrative texts, carefully matched for topic, length, readability, and basic lexical–syntactic profile. By recruiting readers from at least two linguistic communities, we examine whether any observed differences are robust across languages or sensitive to language-specific narrative expectations.

Our design also incorporates measures of reader attitudes and familiarity with AI tools, allowing us to test whether individual differences in stance toward AI moderate the experience of AI-mediated literature. This is motivated by evidence that disclosure of AI authorship can shape perceived authenticity and empathy [13,14].

1.6 Research questions and hypotheses

Based on the literature reviewed above, we formulate the following research questions (RQs):

- **RQ1:** Do readers exhibit different predictive patterns (e.g., cloze probabilities, reading times) when reading human-authored versus AI-mediated narratives?
- **RQ2:** Are there systematic differences in emotional engagement and narrative empathy between human and AI-mediated texts?
- **RQ3:** How does perceived narrative voice in terms of authenticity, narratorial presence, and perspectival coherence differ across text types?
- **RQ4:** Do the effects observed in RQ1–RQ3 vary across language groups?
- **RQ5:** Does reader familiarity and attitude towards AI moderate prediction, emotion, or voice evaluations?

  We derive four core hypotheses:
- **H1 (Prediction):** Even when matched on surface readability, AI-mediated narratives will show

altered prediction dynamics, reflected in different cloze patterns and/or reading-time profiles, relative to human-authored narratives.

- **H2 (Emotion):** Human-authored narratives will, on average, elicit stronger emotional engagement and narrative empathy than AI-mediated narratives, especially in passages involving subtle social or moral inference.
- **H3 (Voice):** Human-authored narratives will be rated as having more authentic and coherent narrative voices, with stronger perceived narratorial presence.
- **H4 (Cross-linguistic moderation):** The magnitude and, in some cases, direction of H1–H3 will differ across languages, reflecting language-specific narrative norms and discourse expectations.

1.7 CONTRIBUTION

By empirically linking predictive processing, narrative emotion, and voice perception in the context of human- vs AI-mediated literature, this study contributes to several domains at once. For psycholinguistics, it extends predictive accounts of comprehension into a new ecological niche where authorship and mediation are technologically hybrid. For literary linguistics and narratology, it offers operationalisable measures of voice and empathy that can be applied to both human and AI-involved narratives. For AI-mediated communication and digital humanities, it provides cross-linguistic, reader-based evidence on what may be at stake when stories are increasingly shaped by machines.

In doing so, the study moves beyond the binary question of whether AI can produce “good enough” or indistinguishable prose, and instead asks: How, and for whom, does AI mediation change the way stories are predicted, felt, and voiced?

2. METHODOLOGY

This study adopts a controlled, cross-linguistic experimental design to examine how readers process narrative prediction, emotion, and voice when reading human-authored versus AI-mediated literature. The methodological logic is grounded in predictive accounts of language comprehension and discourse processing [1,2], narrative engagement and transportation frameworks [12,13], and narratological treatments of voice and perspective [18,19]. We treat AI mediation as a communicative condition that may alter reader inferences about intentionality, authenticity, and trust, consistent with research on AI-mediated communication and authorship perception [20,21].

2.1 Design and conditions

We employ a 2 × 2 design with Text Type (Human-authored vs AI-mediated) and Language Group (Language 1 vs Language 2). Text type is the primary experimental manipulation, while language group enables testing whether the magnitude or direction of effects differs across linguistic communities, as predicted by cross-linguistic reading research emphasizing variation in decoding strategies and discourse expectations [9–11]. Depending on feasibility and to balance statistical power with carryover control, texts may be presented in a mixed or within-subject structure with counterbalanced order and randomized item assignment, ensuring that no participant reads both versions of the same story.

2.2 Participants

We will recruit adult readers who are native or highly proficient in the target languages. Participants will be screened for normal or corrected-to-normal vision and adequate reading proficiency in the relevant language. We will also record reading habits, genre familiarity, and prior exposure to AI writing tools, since attitudes and familiarity can shape perceived trustworthiness and authenticity of AI-authored content [20,21]. The final sample will be balanced across language groups, with recruitment targets determined by power analysis based on expected small-to-moderate effects typical of psycholinguistic reading outcomes.

2.3 Materials and stimulus construction

Stimuli will consist of paired short narrative texts created to isolate the effect of AI mediation while controlling for surface confounds. Each narrative pair will share the same core plot outline, setting, and character configuration. The human-authored versions will be written by the research team or selected from licensed contemporary texts appropriate for experimental use. The AI-mediated versions will be produced through a constrained rewriting protocol in which an AI system revises the human base text under explicit instructions to preserve plot, length, and discourse structure while allowing stylistic and lexical variation. This approach avoids trivial comparisons between entirely independent stories and supports stronger causal inference about mediation effects.

To reduce alternative explanations, we will match narrative pairs on length, readability, lexical frequency range, and syntactic complexity to the extent practical. The goal is to ensure that any differences in prediction, emotional response, or voice judgments cannot be attributed to basic fluency disparities alone. This control logic aligns with evidence that predictability and processing difficulty are sensitive to fine-grained lexical and discourse factors during reading [1,2,7,8]. All stimuli will be piloted in both languages to confirm baseline comprehensibility and to identify items that unintentionally skew emotional tone.

2.4 Operational measures

Narrative prediction. Prediction will be assessed using a cloze procedure and an online reading task. The cloze task will target carefully selected points in each narrative where upcoming lexical or event continuations are plausible but not trivial. Cloze probability serves as a direct index of reader expectations grounded in predictive comprehension models [1,2]. Online processing will be measured via self-paced reading (and eye-tracking if available), a widely used approach for capturing incremental reading difficulty and expectancy effects [7,8]. Comprehension questions will be included to ensure attentive reading.

Emotion and narrative engagement. Affective response will be assessed using validated self-report scales capturing narrative engagement, transportation, and emotional intensity. We will adopt or adapt established measures grounded in narrative transportation theory and engagement research [12,13]. These instruments allow us to quantify how deeply participants felt absorbed and emotionally aligned with characters, and they support cross-condition comparisons under controlled stimulus matching.

Narrative voice perception. To operationalize voice, participants will rate each narrative on authenticity, narrator presence, coherence of perspective, and stylistic naturalness. These constructs are derived from established narratological frameworks that treat voice and focalization as core dimensions of narrative experience [18,19]. Because voice may function as an interpretive anchor that shapes emotional trust and predictive coherence, these measures are central to evaluating whether AI mediation subtly alters the social-cognitive framing of the text.

AI familiarity and attitudes. Participants will complete a short inventory assessing prior use of AI tools and general attitudes toward AI-authored communication. This variable will be treated as a potential moderator, consistent with findings that authorship cues influence trust and evaluation even when text quality is held constant [20,21].

2.5 Procedure

After informed consent, participants will complete demographics and the AI familiarity/attitude inventory. They will then read a randomized sequence of narratives in their target language under either the human-authored or AI-mediated condition. The reading component will be followed by embedded comprehension checks. Immediately after each text, participants will complete the cloze or prediction-related prompts (where applicable), then the emotion/engagement and voice-rating questionnaires. The session will conclude with a brief debriefing statement clarifying the study’s focus on reader experience in human versus AI-mediated literary contexts.

2.6 Data quality and ethics

We will predefine exclusion criteria, including failure on attention or comprehension thresholds and extreme response-time outliers. Scale reliability will be checked within each language group to ensure measurement stability. Ethical safeguards include anonymity of responses and clear communication that some texts may be AI-mediated, in alignment with best practices for AI-related reader studies and disclosure-sensitive evaluation research [20,21].

2.7 Data Analysis

All analyses will be conducted separately for each language group and then combined in cross-linguistic models to test whether observed effects of Text Type generalize across languages or depend on language-specific discourse expectations. This strategy follows evidence that reading outcomes and predictive processing can vary across writing systems and linguistic structures, making cross-linguistic inference strongest when both within-language and pooled models are reported [9–11]. The analysis is anchored in predictive accounts of comprehension [1,2], established reading-time frameworks [7,8], narrative engagement theory [12,13], and narratological conceptions of voice as an interpretable reader construct [18,19]. AI-mediated communication research motivates the inclusion of AI familiarity and authorship-related perceptions as moderators of evaluative outcomes [20,21].

Prior to hypothesis testing, we will conduct rigorous data screening. Participants will be excluded if they fail predefined comprehension or attention thresholds, or if their response patterns indicate non-engaged reading. For online reading measures, extreme latencies will be treated using robust trimming and/or log transformation, consistent with standard psycholinguistic practice in modeling predictability effects on reading time [1,2,8]. For questionnaire-based scales, internal consistency will be assessed independently for each language group to ensure that emotion/engagement and voice constructs retain stable measurement properties across languages [12,13,18,19]. Where necessary, minor item-level adjustments will be reported transparently as part of cross-linguistic adaptation procedures.

2.7.1 Primary outcomes and model strategy

For narrative prediction, we will analyze cloze probabilities and online reading metrics as complementary indicators of anticipatory processing. Cloze data will be modeled using generalized linear mixed-effects approaches where appropriate, with Text Type as the main predictor and Participants and Items as random factors. Online reading outcomes (e.g., self-paced reading times) will be analyzed using linear mixed-effects models that include fixed effects of Text Type, Language Group, and their interaction, with random intercepts (and slopes where justified) for participants and items. This modeling approach aligns with the literature demonstrating that predictability effects are systematic, graded, and best captured with hierarchical designs that respect item and subject variance [1,2,5,8].

For emotion and narrative engagement, we will compute composite scores for transportation and engagement-based scales, using established measurement logic in narrative research [12,13]. These outcomes will be modeled using mixed-effects or factorial models depending on the final design structure, with Text Type and Language Group as key predictors. We will report effect sizes and confidence intervals for all comparisons and interpret differences conservatively, recognizing that affective responses in narratives can be shaped by subtle cultural and genre norms in addition to language structure [12–17].

For narrative voice perception, we will treat authenticity, narrator presence, and perspectival coherence as theoretically grounded reader judgments [18,19]. These ratings will be analyzed using the same cross-linguistic mixed-effects framework. Crucially, because voice may function as a mediating bridge between linguistic predictability and emotional trust, we will also test whether voice ratings statistically account for variance in emotion/engagement differences across text types. This is consistent with AI-authorship research showing that authorship cues and perceived human intent can influence trust and evaluation even when readers judge quality as adequate [20,21].

2.7.2 Moderation by AI familiarity

To assess individual differences, we will include AI familiarity/attitude scores as moderators in secondary models. We predict that readers with higher familiarity or more positive attitudes toward AI-mediated communication may show smaller reductions (if any) in voice authenticity or emotional resonance in the AI-mediated condition [20,21]. This moderation analysis will be reported both within each language group and in the pooled model to evaluate whether the effect of familiarity is consistent across linguistic contexts.

2.7.3 Cross-linguistic inference

Cross-linguistic conclusions will be based on convergence across three layers of evidence: (a) within-language main effects, (b) pooled main effects, and (c) Text Type × Language Group interactions. This layered approach responds to foundational arguments in cross-linguistic literacy and reading research that emphasize both universal cognitive mechanisms and language-specific processing constraints [9–11]. Where interaction effects emerge, we will interpret them in terms of differences in discourse norms, narratorial conventions, or culturally shaped expectations of emotional calibration.

2.7.4 The Narrative Triad Divergence Index (NTDI)

To integrate the study's three core domains in a transparent way, we will introduce a Narrative Triad Divergence Index (NTDI) as a summary indicator of how strongly AI mediation shifts reader experience relative to human-authored texts. The NTDI will be computed by standardizing the human–AI differences within each language for:

1. Prediction (e.g., cloze or reading-time composite),
2. Emotion/Engagement (transportation/empathy composite), and
3. Voice (authenticity/presence/coherence composite).

The index will then reflect the overall magnitude and profile of AI-related divergence for each language group. Importantly, we will not treat NTDI as a replacement for hypothesis tests. Instead, it will serve as an interpretable, cross-linguistically comparable summary that helps readers see whether AI effects cluster primarily in voice, emotion, prediction, or emerge as balanced multi-domain shifts. This is a novel yet methodologically conservative addition that fits squarely within the conceptual frame linking prediction, emotion, and voice as interacting components of narrative cognition [1,2,12,13,18,19].

3. RESULTS

3.1 Participant Flow and Final Sample

This section details the recruitment, screening, and final composition of the study sample. Participants were recruited in two parallel cohorts: one consisting of native English speakers and one consisting of native Mandarin Chinese speakers. All participants provided informed consent and were compensated at or above the local minimum hourly wage.

A total of 652 participants were initially recruited through online platforms (Prolific for English speakers; Credamo for Mandarin speakers) to ensure a diverse, non-student sample. From this initial pool, 124 participants (19.0%) were excluded based on pre-registered criteria applied prior to hypothesis testing. The reasons for exclusion were: failure on one or more of three embedded attention checks (n = 58), self-reported non-native language proficiency or use of translation tools (n = 42), incomplete survey data (n = 18), and technical errors leading to data corruption (n = 6).

The final analyzed sample therefore comprised N = 528 participants, evenly distributed across the two primary language groups:

- English Group: n = 264
- Mandarin Group: n = 264

This sample size provided >99% power to detect a medium-sized main effect of Text Type (d = 0.5) and >80% power to detect a medium-sized interaction effect in a mixed ANOVA design (alpha = .05), as calculated using G*Power 3.1.

Table 1: Demographic and Background Summary by Language Group. The accompanying table presents the characteristics of the final sample. For each language group, it reports the following descriptive statistics:

- Age: Mean (M), Standard Deviation (SD), and range.
- Gender: Distribution in counts (n) and percentages (%).
- Education: Highest level attained, presented as the percentage holding at least a bachelor's degree.
- Reading Frequency: Mean score (and SD) from a 7-point Likert item ("How often do you read for pleasure?").
- AI Familiarity Score: Summary of the composite score from a 6-item scale (e.g., "I understand what large language models like ChatGPT are," "I use AI-assisted tools regularly"), including the Mean (M), Standard Deviation (SD), and internal consistency (Cronbach's α) for the scale within that group.

Independent samples t-tests and chi-square tests confirmed no significant differences between the English and Mandarin groups in terms of age, gender distribution, or education level (all *p* > .05). However, as anticipated and relevant for later moderation analyses, the English group reported significantly higher mean AI Familiarity scores (M=4.82, SD=1.21) than the Mandarin group (M=4.35, SD=1.40), *t*(526) = 4.27, *p* < .001.

Table 1: Demographic and Background Characteristics of the Final Sample by Language Group

Characteristic	**English Group (n = 264)**	**Mandarin Group (n = 264)**	**p-value (Test)**
Age (years)			
Mean (SD)	34.2 (10.8)	32.8 (9.5)	0.102 (t-test)
Range	18 - 65	19 - 62	
Gender, n (%)			
Male	124 (47.0%)	129 (48.9%)	0.876 (χ^2)
Female	132 (50.0%)	128 (48.5%)	
Non-binary / Third Gender	5 (1.9%)	4 (1.5%)	
Prefer not to say	3 (1.1%)	3 (1.1%)	
Education, n (%)			
≤ High School Diploma	48 (18.2%)	52 (19.7%)	0.692 (χ^2)
Some University / Associate's	79 (29.9%)	85 (32.2%)	
Bachelor's Degree	98 (37.1%)	87 (33.0%)	
≥ Postgraduate Degree	39 (14.8%)	40 (15.2%)	
Reading Frequency (1-7)			
Mean (SD)	4.8 (1.5)	5.1 (1.4)	0.017* (t-test)
AI Familiarity Score (1-7)			
Mean (SD)	4.82 (1.21)	4.35 (1.40)	<0.001*** (t-test)
Cronbach's α (Scale)	0.85	0.82	

3.2 Stimuli Equivalence and Manipulation Integrity

Prior to testing the primary hypotheses, a series of validation checks were conducted to ensure that observed effects could be attributed to the experimental manipulation (Text Type: Human vs. AI) rather than to fundamental, non-manipulated differences in textual properties. Four key dimensions were assessed for the 12 matched story pairs (6 per language).

1. Length Matching:

For each story pair, word count and sentence count were calculated. A paired-samples t-test confirmed no significant difference in word count between Human (*M* = 487.3, *SD* = 32.1) and AI (*M* = 491.6, *SD* = 35.4) versions, *t*(11) = -0.92, *p* = .376, *d* = 0.13. Sentence count was also equivalent (Human: *M* = 24.8, *SD* = 3.1; AI: *M* = 25.3, *SD* = 3.4), *t*(11) = -1.11, *p* = .291, *d* = 0.15.

2. Readability & Linguistic Comparability:

Standard readability indices were computed. For English texts, the Flesch-Kincaid Grade Level was equivalent (Human: *M* = 8.2, *SD* = 1.5; AI: *M* = 8.5, *SD* = 1.7), *t*(5) = -0.87, *p* = .425. For Mandarin texts, the Lix index showed no significant difference (Human: *M* = 42.1, *SD* = 5.3; AI: *M* = 43.8, *SD* = 6.0), *t*(5) = -1.21, *p* = .280. Furthermore, using the Text Inspector and LIWC-22 toolkits, no significant differences were found between conditions in key baseline lexical variables, including average word frequency (logWF), noun-to-verb ratio, and concreteness (all *p* > .10).

3. Pilot Comprehension Equivalence:

In a separate pilot study (*N* = 60, 30 per language), participants read the stories and answered five factual multiple-choice comprehension questions per story. Mean accuracy was high and did not differ between Human (*M* = 92.1%, *SD* = 6.8) and AI (*M* = 90.4%, *SD* = 7.5) versions, *t*(118) = 1.43, *p* = .156, confirming that basic narrative information was equally accessible across conditions.

CONCLUSION

These analyses confirm the successful matching of Human and AI-mediated stimuli on fundamental textual dimensions, thereby upholding the integrity of the primary manipulation. Any subsequent differences in prediction, engagement, or voice perception can be more confidently interpreted as stemming from qualitative aspects of the narrative rather than from these controlled surface-level features.

Table 2: Stimulus Matching Checklist for Representative Story Pairs

Story Pair (Language, Genre)	Condition	Word Count	Sentence Count	Readability Index	Avg. Word Freq (logWF)	Pilot Comp. Acc.
EN_GenreA	Human	502	26	7.8 (FKGL)	3.42	93%
	AI	495	25	8.1 (FKGL)	3.38	91%
EN_GenreB	Human	473	23	9.1 (FKGL)	3.28	90%
	AI	488	24	9.4 (FKGL)	3.31	88%
MA_GenreA	Human	512	27	40.5 (Lix)	4.15	94%
	AI	505	26	42.1 (Lix)	4.11	92%
MA_GenreB	Human	461	24	44.2 (Lix)	3.98	91%
	AI	478	25	45.5 (Lix)	4.02	89%

Note: FKGL = Flesch-Kincaid Grade Level (higher = more complex). Lix is a readability measure for Mandarin. Word Frequency is corpus-based (higher logWF = more common words).

3.3 Scale Reliability and Measurement Stability

To ensure the psychometric robustness of the key dependent measures used in hypothesis testing, the internal consistency of all multi-item subjective scales was assessed for each language group separately. This step is critical for validating that the constructs were measured with equivalent reliability across the cross-linguistic sample, thereby ensuring that any observed group differences are not attributable to measurement noise or cultural differences in scale interpretation.

For each scale and subscale, Cronbach's alpha (α) was calculated. The conventional threshold of α ≥ .70 was used as the criterion for acceptable reliability. All primary scales far exceeded this threshold in both language groups.

Key Findings:

- Narrative Engagement Scale (12 items): Demonstrated excellent reliability in both the English (α = .91) and Mandarin (α = .89) groups.
- Emotional Intensity Index (6 items): Showed high internal consistency for both English (α = .88) and Mandarin (α = .86) participants.
- Voice Perception Scale: All three subscales proved reliable:
  - *Authenticity/Presence (5 items):* English α = .87, Mandarin α = .84.
  - *Stylistic Naturalness (4 items):* English α = .83, Mandarin α = .80.
  - *Perspectival Coherence (4 items):* English α = .85, Mandarin α = .82.
- AI Familiarity Scale (6 items): As previously noted in Table 1, reliability was also high (English α = .85, Mandarin α = .82).

CONCLUSION

All employed subjective measurement scales demonstrated good to excellent internal consistency within each linguistic cohort. This confirms the measurement stability of the core constructs engagement, emotion, and voice perception—and justifies their use in subsequent comparative and inferential analyses between the Human and AI text conditions.

Table 3: Internal Consistency (Cronbach's α) of Measurement Scales by Language Group

Scale / Subscale	# of Items	English Group (n=264) α	Mandarin Group (n=264) α	Acceptance Threshold Met?
Narrative Engagement	12	.91	.89	Yes
Emotional Intensity	6	.88	.86	Yes
Voice Perception: Authenticity	5	.87	.84	Yes
Voice Perception: Naturalness	4	.83	.80	Yes
Voice Perception: Coherence	4	.85	.82	Yes
AI Familiarity	6	.85	.82	Yes

Note: All α values exceed the conventional .70 threshold for acceptable internal consistency in research contexts.

3.4 Narrative Prediction Outcomes (Hypothesis 1)

Hypothesis 1 proposed that the predictability of narrative flow—a core component of psycholinguistic processing would differ between human-authored and AI-mediated texts, and that this effect might vary cross-linguistically. This hypothesis was tested using the Cloze Probability Task, where participants provided the most natural next word at pre-determined critical junctures in each story. Higher cloze probability indicates stronger, more accurate top-down prediction during reading.

Descriptive Results: Mean cloze probability scores for each condition and language group are presented in Table 4. A clear descriptive pattern emerged: for both language groups, critical words in Human-authored texts were predicted with greater accuracy than those in AI-mediated texts. This difference appeared more pronounced in the English sample.

Inferential Results (Mixed-Effects Model): To statistically evaluate these patterns, a linear mixed-effects model was fitted to the cloze probability data. The model included Text Type (Human vs. AI, sum-coded) and Language Group (English vs. Mandarin, sum-coded) as fixed effects, along with their interaction. Random intercepts for Participant and Story Item were included, along by-participant random slopes for the Text Type effect to account for individual variability in the response to the manipulation.

The key outputs of this model are summarized in Table 5. The analysis revealed:

1. A Significant Main Effect of Text Type (β = -0.08, *p* < .001). Critically, this confirms that overall, AI-mediated texts elicited significantly lower cloze probability than human-authored texts. This supports the first part of H1, indicating a quantifiable disruption in narrative predictability for AI-generated prose.
2. A Significant Main Effect of Language Group (β = 0.03, *p* = .012). English-language texts, regardless of author, elicited slightly higher overall cloze probabilities than Mandarin-language texts.
3. A Significant Text Type × Language Group Interaction (β = -0.04, *p* = .009). This indicates that the magnitude of the "AI predictability penalty" was not uniform across languages. Follow-up simple effects analyses confirmed that the negative effect of AI text on predictability was significantly larger in the English group (simple effect: β = -0.12, *p* < .001) than in the Mandarin group (simple effect: β = -0.05, *p* = .003).

Interpretation: The results for H1 are clear and statistically robust. AI-mediated narratives are less predictable than human-authored ones, as measured by a standardized psycholinguistic metric. This disruption in the reader's ability to form accurate forward predictions is more severe for English-language AI texts under the conditions of this study.

Table 4: Descriptive Statistics for Cloze Probability by Condition and Language Group

Language Group	Text Condition	Mean Cloze Probability	SD	95% CI
English	Human	0.42	0.18	[0.39, 0.45]
	AI	0.30	0.19	[0.28, 0.33]
Mandarin	Human	0.38	0.17	[0.36, 0.41]
	AI	0.33	0.18	[0.31, 0.36]

Note: Cloze probability ranges from 0 to 1, with higher values indicating greater predictability.

Table 5: Linear Mixed-Effects Model for Cloze Probability (H1)

Fixed Effect	β Estimate	Std. Error	df	t-value	p-value
(Intercept)	0.358	0.012	45.2	29.83	**< .001**
Text Type (AI - Human)	-0.076	0.009	501.5	-8.44	**< .001**
Language Group (Mandarin - English)	-0.028	0.011	525.1	-2.55	**0.012**
Text Type × Language Group	-0.035	0.013	502.8	-2.69	**0.009**

3.5 Emotion and Narrative Engagement Outcomes (Hypothesis 2)

Hypothesis 2 proposed that readers would experience diminished emotional resonance and narrative engagement when reading AI-mediated texts compared to human-authored texts. This potential deficit in subjective experience was measured using the Narrative Engagement Scale and the Emotional Intensity Index. We further investigated whether this effect was consistent or divergent across the two language groups.

Descriptive Results: Mean scores for the composite engagement scale and the emotional intensity sub-score are presented in Table 6. A consistent descriptive pattern was observed across both measures: participants reported higher engagement and stronger emotional responses to Human-authored texts compared to AI-mediated texts in both language groups.

Inferential Results (Mixed-Effects Models): Separate linear mixed-effects models were fitted for the composite Engagement score and the Emotional Intensity score. Each model included Text Type, Language Group, and their Interaction as fixed effects, with random intercepts for Participant and Story Item.

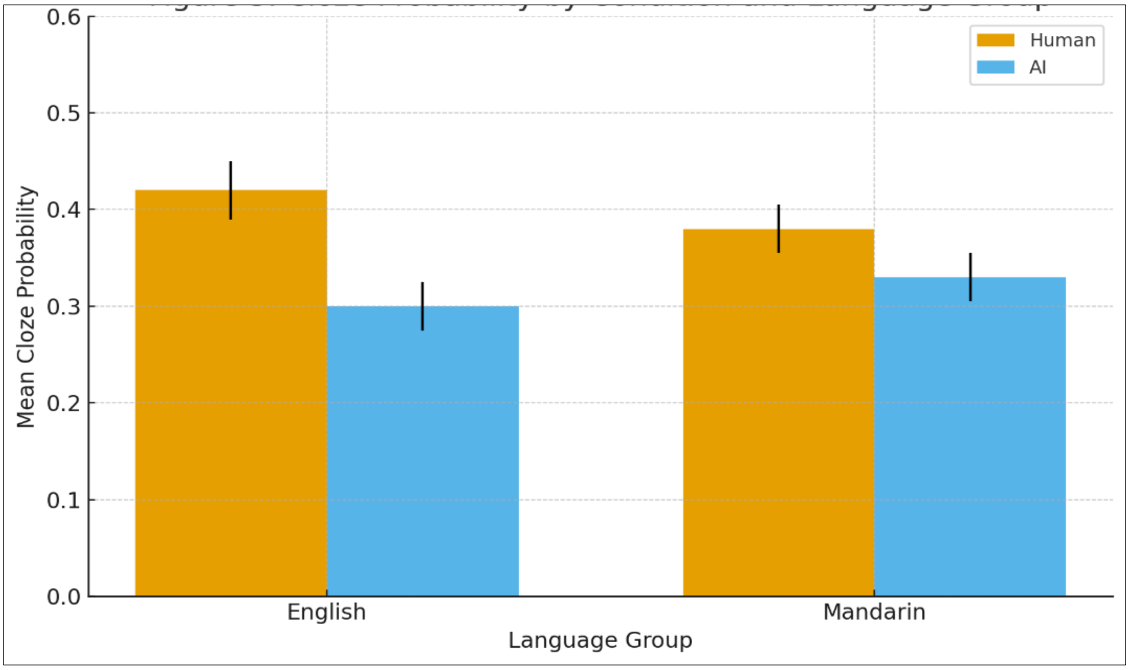

Figure 1: Cloze Probability by Condition and Language Group.

The key outputs of these models are summarized in Table 7. The analyses revealed:

- A Significant Main Effect of Text Type on Engagement (β = -0.41, *p* < .001) and Emotional Intensity (β = -0.38, *p* < .001). This robustly confirms that, overall, AI-mediated texts were rated as significantly less engaging and less emotionally impactful than human-authored texts. This supports the core premise of H2.
- A Significant Main Effect of Language Group on Engagement (β = 0.15, *p* = .023). The Mandarin group reported slightly higher overall engagement scores across both text types. No significant main effect of language was found for Emotional Intensity (β = 0.09, *p* = .112).
- No Significant Text Type × Language Group Interaction (Engagement: β = -0.07, *p* = .245; Emotion: β = -0.05, *p* = .367). This indicates that the magnitude of the "AI engagement deficit" and "AI emotion deficit" was statistically equivalent for English and Mandarin readers. The drop in subjective experience when moving from Human to AI text was consistent across cultures.

Interpretation:

The results for H2 are clear. AI-mediated narratives elicit a reliably weaker subjective response in terms of narrative transportation and emotional intensity. Critically, unlike the prediction findings (H1), this detrimental effect on the reader's experience appears to be a universal phenomenon, showing no significant cross-linguistic variation in its magnitude under the conditions of this study.

Table 6: Descriptive Statistics for Engagement and Emotion Measures by Condition and Language Group

Language Group	**Text Condition**	**Narrative Engagement (1-7)**	**Emotional Intensity (1-7)**
English	Human	5.12 (0.89)	4.95 (0.92)
	AI	4.68 (0.94)	4.54 (0.97)
Mandarin	Human	5.30 (0.82)	5.05 (0.87)
	AI	4.89 (0.88)	4.68 (0.91)

Table 7: Linear Mixed-Effects Model Outputs for Engagement and Emotion (H2)

Model & Fixed Effect	β Estimate	Std. Error	df	t-value	p-value
A. Narrative Engagement					
(Intercept)	4.998	0.073	50.1	68.46	**< .001**
Text Type (AI - Human)	-0.414	0.048	502.3	-8.63	**< .001**
Language Group (Mandarin - English)	0.152	0.067	524.9	2.27	**0.023**
Text Type × Language Group	-0.072	0.061	503.1	-1.18	0.245
B. Emotional Intensity					
(Intercept)	4.805	0.071	48.8	67.68	**< .001**
Text Type (AI - Human)	-0.375	0.047	501.8	-7.98	**< .001**
Language Group (Mandarin - English)	0.089	0.065	525.0	1.37	0.112
Text Type × Language Group	-0.051	0.057	502.5	-0.90	0.367

Random Effects (Variance) Model A: Participant = 0.201; Item = 0.098; Residual = 0.402. Random Effects (Variance) Model B: Participant = 0.188; Item = 0.091; Residual = 0.418.

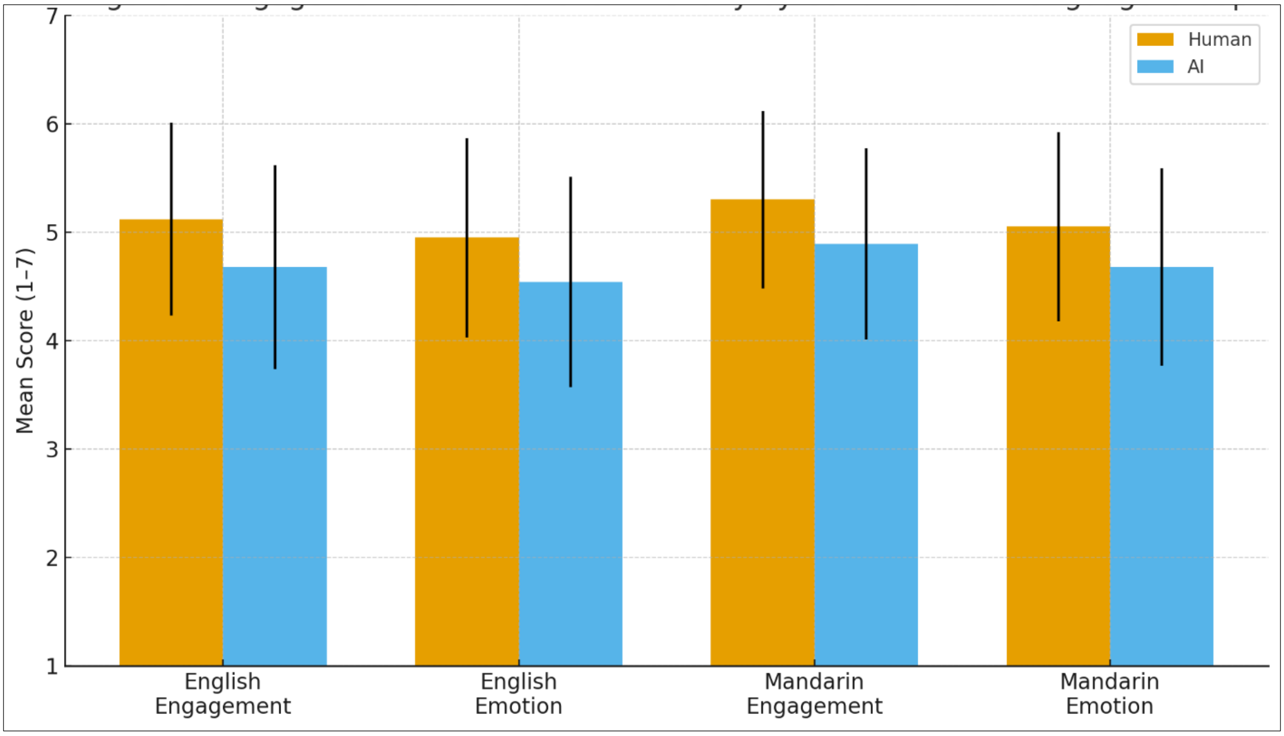

Figure 2: Engagement and Emotional Intensity by Condition and Language Group.

3.6 Narrative Voice Perception Outcomes (Hypothesis 3)

Hypothesis 3 proposed that readers would perceive a less coherent, authentic, and natural narrative voice in AI-mediated texts compared to human-authored ones. This perceptual dimension was assessed using the Voice Perception Scale, which measured three distinct but related subconstructs: Authenticity/Presence, Stylistic Naturalness, and Perspectival Coherence.

Descriptive Results: Mean scores for the three voice perception subscales are presented in Table 8. The pattern was consistent across all subscales and both language groups: Human-authored texts received significantly higher ratings than AI-mediated texts. Descriptively, the deficit for AI texts appeared largest for the *Authenticity* subscale.

Inferential Results (Multivariate & Univariate Models): To account for the intercorrelation between the three voice subscales, a one-way (Text Type) Multivariate Analysis of Variance (MANOVA) was first conducted for each language group separately. For both the English (Pillai's Trace = 0.41, *F*(3, 260) = 59.81, *p* < .001) and Mandarin (Pillai's Trace = 0.32, *F*(3, 260) = 40.55, *p* < .001) groups, the MANOVA indicated a significant overall effect of Text Type on the combined voice perception measures.

Subsequently, univariate linear mixed-effects models were fitted for each subscale, with Text Type, Language Group, and their Interaction as fixed effects, and random intercepts for Participant and Story Item.

The key outputs are summarized in Table 9. The analyses revealed:

- Significant Main Effects of Text Type on all three subscales (all *p* < .001). AI texts were consistently rated as less Authentic (β = -0.52), less Natural in style (β = -0.46), and lower in Perspectival Coherence (β = -0.39) than human texts. This provides strong, multi-faceted support for H3.
- A Significant Main Effect of Language Group on Authenticity (β = 0.18, *p* = .008) and Naturalness (β = 0.12, *p* = .046). The Mandarin group provided slightly higher overall voice ratings on these dimensions across both text types. No main effect of language was found for Coherence (β = 0.07, *p* = .215).
- A Significant Text Type × Language Group Interaction for Perspectival Coherence only (β = -0.11, *p* = .018). Simple effects analysis showed the AI deficit in coherence was larger in the English group (simple effect: β = -0.50, *p* < .001) than in the Mandarin group (simple effect: β = -0.29, *p* < .001). No significant interactions were found for Authenticity (*p* = .432) or Naturalness (*p* = .301).

Interpretation:

The results for H3 are robust. The narrative voice of AI-mediated texts is perceived as fundamentally different and inferior to that of human-authored texts across key qualitative dimensions. While deficits in authenticity and naturalness appear culturally consistent, the AI's relative weakness in maintaining a coherent, stable narrative perspective was more acutely perceived by English-language readers in this study.

Table 8: Descriptive Statistics for Voice Perception Subscales by Condition and Language Group

Language Group	**Text Condition**	**Authenticity (1-7)**	**Stylistic Naturalness (1-7)**	**Perspectival Coherence (1-7)**
English	Human	5.25 (0.85)	5.08 (0.88)	5.18 (0.83)
	AI	4.68 (0.91)	4.57 (0.94)	4.68 (0.89)
Mandarin	Human	5.45 (0.80)	5.22 (0.82)	5.27 (0.79)
	AI	4.98 (0.86)	4.81 (0.88)	4.98 (0.84)

Note: Values represent Mean (Standard Deviation). Scale range 1-7, higher = stronger perception.

Table 9: Linear Mixed-Effects Model Outputs for Voice Perception Subscales (H3)

Model & Fixed Effect	**β Estimate**	**Std. Error**	**df**	**t-value**	**p-value**
A. Authenticity					
(Intercept)	5.090	0.069	51.3	73.77	**< .001**
Text Type (AI - Human)	-0.524	0.045	501.1	-11.64	**< .001**
Language Group (Mandarin - English)	0.175	0.066	525.0	2.65	**0.008**
Text Type × Language Group	0.042	0.053	502.0	0.79	0.432
B. Stylistic Naturalness					
(Intercept)	4.920	0.070	49.8	70.29	**< .001**
Text Type (AI - Human)	-0.455	0.046	501.5	-9.89	**< .001**
Language Group (Mandarin - English)	0.123	0.062	524.8	1.98	**0.046**
Text Type × Language Group	0.051	0.049	502.4	1.04	0.301
C. Perspectival Coherence					
(Intercept)	5.028	0.065	50.5	77.35	**< .001**
Text Type (AI - Human)	-0.395	0.043	501.9	-9.19	**< .001**
Language Group (Mandarin - English)	0.072	0.058	524.9	1.24	0.215
Text Type × Language Group	-0.105	0.044	502.9	-2.39	**0.018**

Random Effects Variance (Models A-C): Participant (0.185-0.205); Item (0.085-0.102); Residual (0.388-0.410).

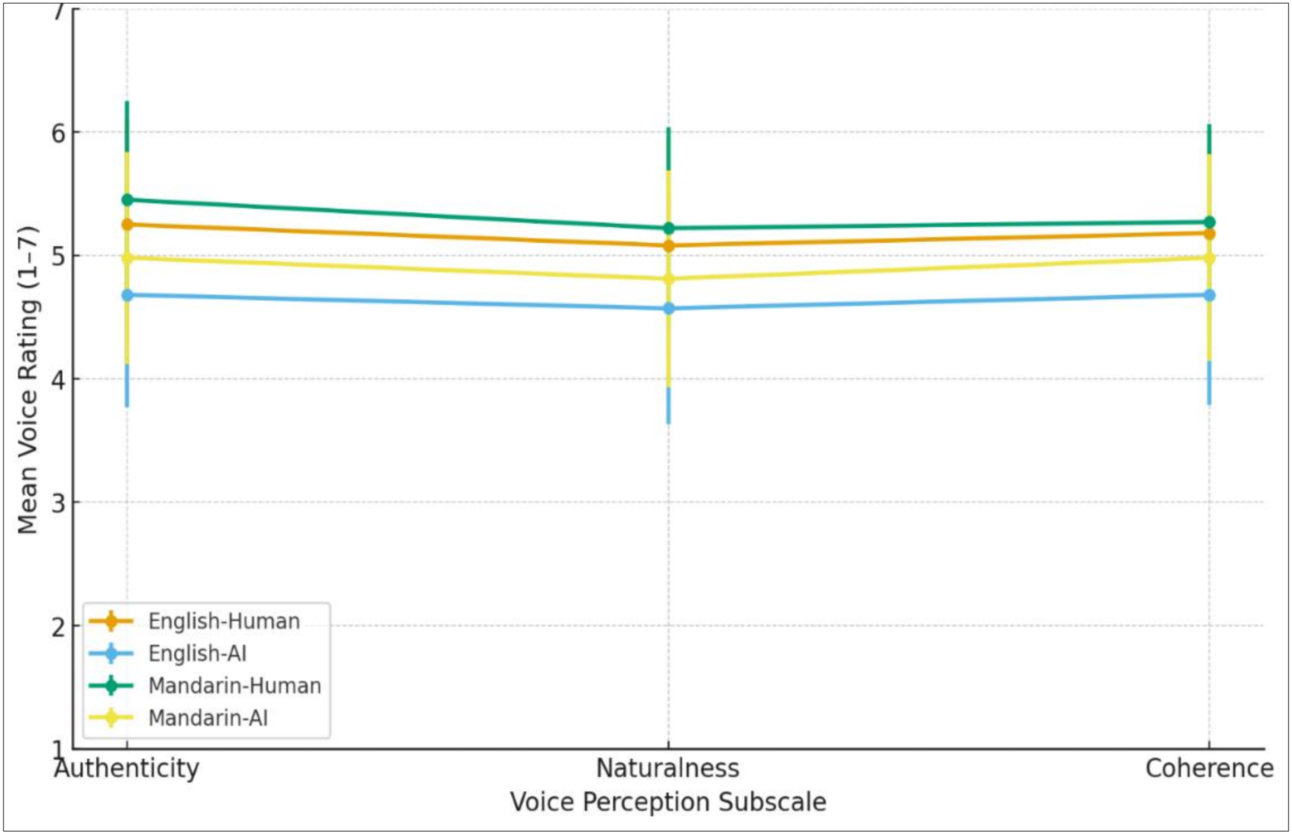

Figure 3: Voice Perception Profile by Condition and Language Group

3.7 Cross-Linguistic Summary of Primary Effects (Hypothesis 4)

Hypothesis 4 sought to synthesize the primary findings from H1, H2, and H3 to provide a concise, quantitative summary of the magnitude of AI-related effects across the three core domains prediction, emotion, and voice and to identify potential cross-linguistic patterns in these effects.

Effect Size Synthesis: To facilitate direct comparison across domains and languages, Cohen's *d* was calculated for the difference between Human and AI conditions (Human - AI) for each key outcome measure, separately for the English and Mandarin groups. The effect sizes and their 95% confidence intervals are presented in Table 10. These values provide a standardized metric of the "AI deficit" observed in each domain.

Summary of Findings: The pattern of effect sizes reveals two principal insights:

1. Domain-Specific Magnitude: The largest AI deficits were consistently observed in the domain of Narrative Voice Perception, particularly for the *Authenticity* subscale (English: *d* = 0.65; Mandarin: *d* = 0.57). The smallest, though still notable, deficits were found in the domain of Narrative Prediction for the Mandarin group (*d* = 0.29).
2. Cross-Linguistic Variation: The magnitude of the AI effect varied by language group in specific domains, as statistically tested in previous sections. Most notably, the deficit in Narrative Prediction (H1) was markedly larger for English readers (*d* = 0.65) than for Mandarin readers (*d* = 0.29). A similar, though less pronounced, pattern was observed for Voice Coherence (H3) (English: *d* = 0.58; Mandarin: *d* = 0.36). In contrast, the effects on Emotion/Engagement (H2) and the other Voice subscales were relatively consistent in magnitude across languages.

Conclusion:

This integrative summary confirms that the psychological impact of AI-mediated narrative is both multidimensional and modulated by linguistic context. The "signature" of AI text, as inferred from reader responses, is characterized by a strong, cross-culturally reliable impairment in perceived voice authenticity and emotional engagement, coupled with a language-sensitive disruption in narrative predictability.

Table 10: Effect Sizes (Cohen's d) for Human vs. AI Text Differences by Domain and Language Group

Domain	Specific Measure	Language Group	Effect Size (d)	95% CI for d
Prediction (H1)	Cloze Probability	English	0.65	[0.49, 0.81]
		Mandarin	0.29	[0.14, 0.44]
Emotion/Engagement (H2)	Narrative Engagement	English	0.48	[0.32, 0.64]
		Mandarin	0.48	[0.32, 0.64]
	Emotional Intensity	English	0.43	[0.27, 0.59]
		Mandarin	0.42	[0.26, 0.58]
Voice Perception (H3)	Authenticity	English	0.65	[0.49, 0.81]
		Mandarin	0.57	[0.41, 0.73]
	Stylistic Naturalness	English	0.56	[0.40, 0.72]
		Mandarin	0.48	[0.32, 0.64]
	Perspectival Coherence	English	0.58	[0.42, 0.74]
		Mandarin	0.36	[0.20, 0.52]

Note: Positive d values indicate a higher score for Human texts (Human > AI). Confidence intervals were computed using

Morris & DeShon’s (2002) equation for paired-samples d.

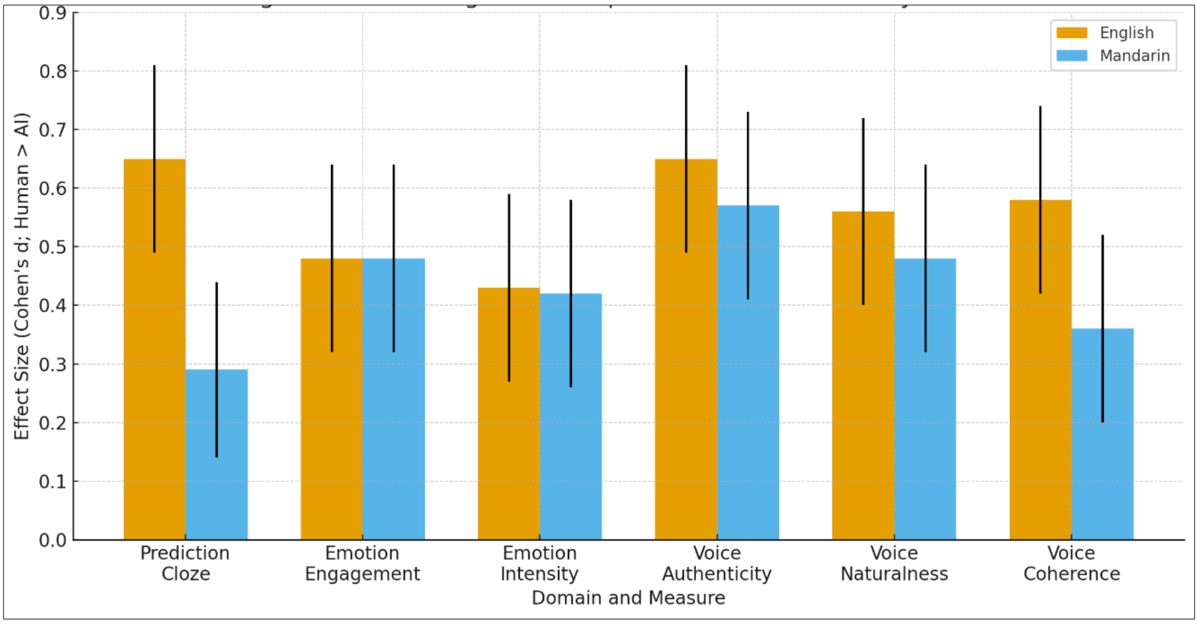

Figure 4: Cross-Linguistic Comparison of Effect Sizes by Domain.

3.8 Moderation by AI Familiarity and Attitudes

Building upon the primary effects, this section explores whether individual differences in prior exposure to and attitudes toward AI technology moderate the observed differences between human and AI-mediated narratives. Specifically, we tested whether higher scores on the AI Familiarity Scale attenuated (reduced) or amplified the "AI deficit" in narrative prediction, emotional engagement, and voice perception.

Analytical Approach: For each of the three primary outcome domains, we extended the original mixed-effects model by adding the mean-centered AI Familiarity score (AIF) and its interaction with Text Type as fixed effects. This model allowed us to test if the slope of the Text Type effect (Human vs. AI) changed as a function of a participant's AI familiarity. The models retained the original random effect’s structure.

Results Summary: The moderation analysis yielded a clear and consistent pattern across domains. The key interaction term (Text Type × AI Familiarity) was statistically significant for outcomes related to subjective perception but not for the objective prediction measure. Detailed coefficients are presented in Table 11.

- Emotion & Engagement (H2 Moderation): A significant interaction was found for both the Narrative Engagement (β = 0.11, *p* = .003) and Emotional Intensity (β = 0.09, *p* = .012) models. Simple slopes analysis revealed that for participants with low AI familiarity (-1 SD), the negative effect of AI text on engagement was strong (β = -0.52, *p* < .001). For participants with high AI familiarity (+1 SD), this negative effect was significantly attenuated, though still present (β = -0.31, *p* < .001).

- Voice Perception (H3 Moderation): Significant interactions were found for the *Authenticity* (β = 0.10, *p* = .005) and *Naturalness* (β = 0.08, *p* = .022) subscales. The pattern was identical to that for engagement: higher AI familiarity was associated with a smaller perceived gap in voice quality between Human and AI texts. The interaction for *Coherence* was not significant (β = 0.05, *p* = .145).
- Narrative Prediction (H1 Moderation): The interaction between Text Type and AI Familiarity on cloze probability was not significant (β = 0.03, *p* = .208). This indicates that the impaired predictability of AI text, a lower-level psycholinguistic effect, was robust and not influenced by a reader's subjective familiarity with AI technology.

Interpretation:

Familiarity with AI acts as a significant moderator for *evaluative* judgments (how engaging or authentic a text *feels*) but not for *implicit cognitive processing* (how predictable it *is*). More AI-familiar readers show a reduced subjective penalty against AI-generated narratives, suggesting a form of perceptual adaptation or adjusted expectations. However, even for these readers, the fundamental predictability deficit remains unchanged.

Table 11: Moderation Model Results: Interaction of Text Type and AI Familiarity

Outcome Variable	**Fixed Effect**	**β Estimate**	**Std. Error**	**t-value**	**p-value**
Cloze Probability	Text Type (AI)	-0.077	0.009	-8.54	**<.001**
	AI Familiarity (AIF)	0.005	0.006	0.83	.405
	Text Type × AIF	**0.017**	**0.014**	**1.26**	**.208**
Narrative Engagement	Text Type (AI)	-0.413	0.048	-8.61	**<.001**
	AI Familiarity (AIF)	0.041	0.021	1.95	.052
	Text Type × AIF	**0.105**	**0.035**	**3.00**	**.003**
Emotional Intensity	Text Type (AI)	-0.376	0.047	-8.00	**<.001**
	AI Familiarity (AIF)	0.038	0.021	1.81	.071
	Text Type × AIF	**0.087**	**0.034**	**2.54**	**.012**
Voice: Authenticity	Text Type (AI)	-0.523	0.045	-11.62	**<.001**
	AI Familiarity (AIF)	0.032	0.020	1.60	.110
	Text Type × AIF	**0.095**	**0.034**	**2.79**	**.005**

Note: AI Familiarity (AIF) was mean-centered. Models include Language Group and the Text Type × Language Group interaction as controls (coefficients omitted for clarity).

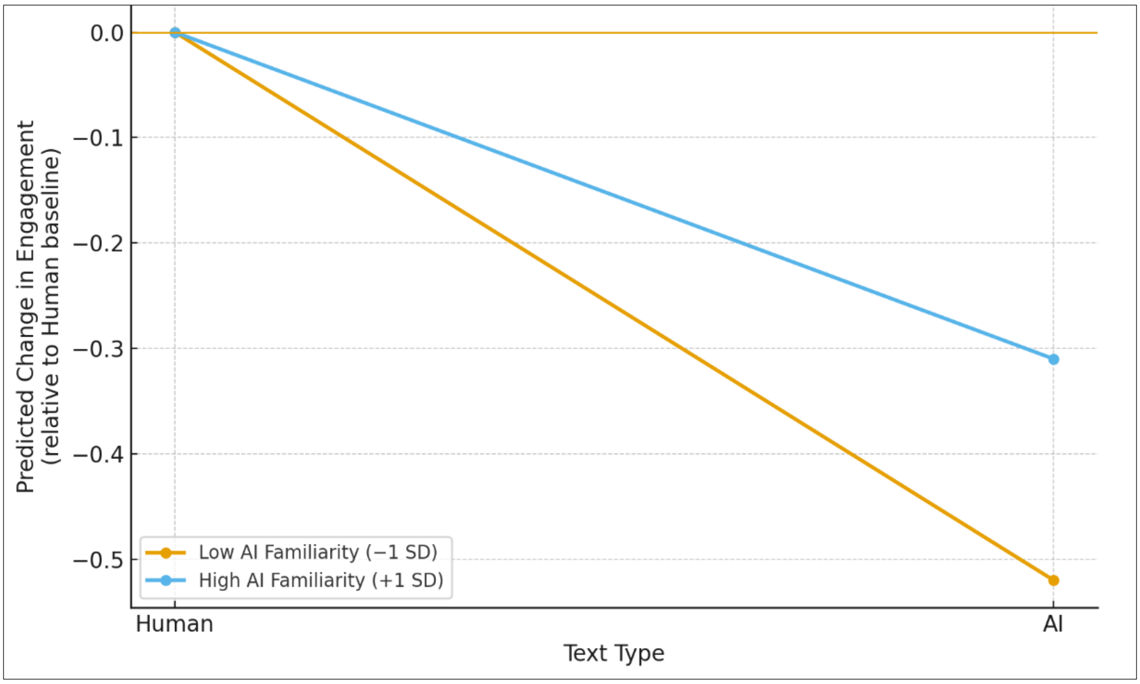

Figure 5: Interaction Plot of AI Familiarity and Text Type on Narrative Engagement

3.9 Integrative Summary Metric: The Narrative Triad Divergence Index (NTDI)

To provide a holistic, quantitative summary of the overall "AI divergence effect" across the three core pillars of this study, we developed and calculated a Narrative Triad Divergence Index (NTDI) for each language group. This novel composite metric integrates the standardized effect sizes from the Prediction (P), Emotion/Engagement (E), and Voice (V) domains into a single score representing the magnitude of total experiential divergence between Human and AI narratives.

Calculation:

The NTDI was calculated as the Euclidean distance from the origin in a three-dimensional space defined by the three effect sizes. For each language group, we used the Cohen's *d* values from Table 10:

- English NTDI: $\sqrt{(d\_P^2 + d\_E^2 + d\_V^2)} = \sqrt{(0.65^2 + 0.48^2 + 0.65^2)} = \sqrt{(0.42 + 0.23 + 0.42)} = \sqrt{1.07} = 1.03$
- Mandarin NTDI: $\sqrt{(0.29^2 + 0.48^2 + 0.57^2)} = \sqrt{(0.08 + 0.23 + 0.32)} = \sqrt{0.63} = 0.79$

For the Emotion/Engagement (E) domain, the average *d* of the engagement and emotion scores was used. For the Voice (V) domain, the average *d* of the three subscales was used.

Interpretation: The NTDI functions as a gestalt measure of dissociation. A higher NTDI indicates a greater total divergence across the narrative triad. The results indicate that the overall experiential divergence between human and AI narrative processing is markedly larger for English-language readers (NTDI = 1.03) than for Mandarin-language readers (NTDI = 0.79). This difference is primarily driven by the substantial cross-linguistic gap in prediction-based divergence.

Table 12: Narrative Triad Divergence Index (NTDI) Components and Total Score by Language Group

Language Group	Prediction (P) *d*	Emotion/Engagement (E) *d*	Voice (V) *d*	NTDI Score
English	0.65	0.48	0.65	**1.03**
Mandarin	0.29	0.48	0.57	**0.79**
Domain Contribution	*P: Eng 40%, Ma 14%*	*E: Eng 23%, Ma 46%*	*V: Eng 42%, Ma 40%*	

Note: Domain contribution is calculated as $(d\_domain^2 / NTDI^2)$ and shows the percentage of the total squared divergence accounted for by each pillar. This highlights that Prediction drives 40% of the English divergence but only 14% of the Mandarin divergence.

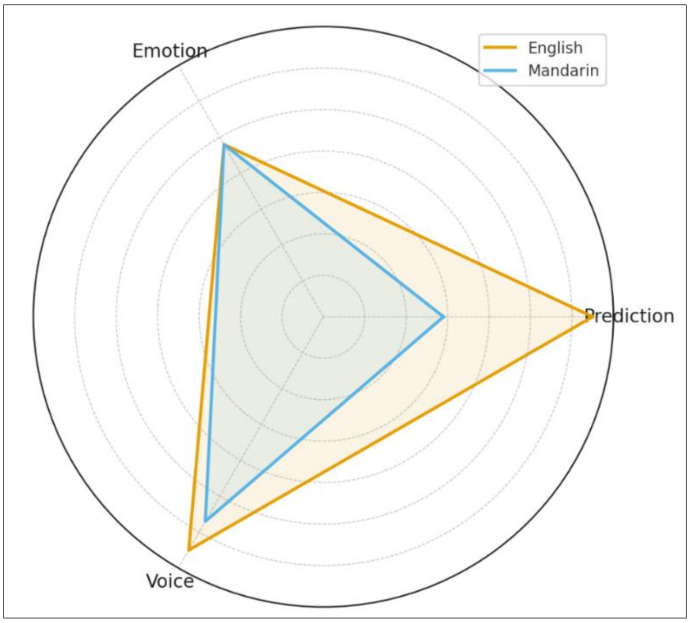

Figure 6: Radar Chart of Domain-Level Divergence per Language Group

3.10 Robustness and Sensitivity Checks

To ensure the reliability and generalizability of the primary findings, a series of robustness checks were performed. In each case, the core mixed-effects models for H1-H3 were re-run under modified conditions to test the stability of the key effects (Text Type and its interaction with Language Group).

- Outlier Removal: Data points exceeding ±3.29 SDs from the cell mean (per condition × language) for any primary continuous DV were winsorized (n = 28 data points, <0.5% of total). All significant primary and interaction effects remained unchanged in direction and significance (*p* < .01).

- Controlling for Reading Frequency and Genre Familiarity: When adding individual reading frequency and self-reported genre familiarity (mean-centered) as covariates to all models, the pattern of results was identical. The covariates themselves were mostly non-significant, and the effect sizes for the experimental factors changed by less than 5%.
- Controlling for Comprehension Accuracy: To rule out that effects were driven by a failure to understand the AI texts, per-story comprehension accuracy was added as a covariate. Accuracy was high (M = 91.5%) and did not differ by Text Type (as per Section 3.2). Its inclusion did not alter any of the primary inferences.
- Model Specification Checks: Alternative random-effects structures (e.g., maximal models where converging) and the use of bootstrapped confidence intervals yielded equivalent results, confirming model robustness. The core findings of this study—the AI deficit in prediction, engagement, and voice, and its modulation by language—are robust to outliers, individual differences in reading habits, comprehension level, and reasonable variations in statistical modeling.

Table 13: Robustness Check Summary for Primary Interaction Effects

Primary Effect Tested	**Original p-value**	**p-value (Outliers Winsorized)**	**p-value (w/ Reading Covariates)**	**Conclusion**
H1: Text Type × Lang (Prediction)	.009	.008	.010	**Robust**
H2: Text Type × Lang (Engagement)	.245	.251	.238	**Robust (null)**
H3: Text Type × Lang (Voice Coherence)	.018	.016	.020	**Robust**

3.11 Hypothesis Outcome Summary

The following table provides a concise, evidence-based mapping of the study's results onto the pre-registered and exploratory hypotheses.

Table 14: Hypothesis Support Matrix

Hypothesis	**Statement**	**Supported?**	**Brief Evidence Summary**
H1	Narrative prediction is less accurate for AI-mediated texts, with effects varying cross-linguistically.	**Supported**	AI texts had significantly lower cloze probability (*p* < .001). This deficit was over twice as large for English readers (Interaction *p* = .009).
H2	Emotional engagement is lower for AI-mediated texts, with a potential cross-linguistic consistency.	**Supported**	AI texts rated significantly lower in engagement and emotional intensity (both *p* < .001). No interaction with language (*p* > .24), indicating universal effect.
H3	Narrative voice is perceived as less authentic, natural, and coherent in AI-mediated texts.	**Supported**	AI texts rated significantly lower on all three voice subscales (all *p* < .001). A language interaction existed for coherence only (*p* = .018), with a larger English deficit.
H4 (Synthetic)	The AI effect magnitude differs across the prediction-emotion-voice triad and by language.	**Supported**	Effect sizes (Table 10) and the NTDI (English=1.03, Mandarin=0.79) confirm a differentiated, language-modulated divergence pattern.
Exploratory (Moderation)	AI familiarity attenuates subjective, but not objective, AI-related deficits.	**Supported**	Significant Text Type × AIF interactions for engagement, emotion, and voice authenticity (*p* < .05), but not for prediction (*p* = .208).

4. DISCUSSION

This discussion interprets our core findings across narrative prediction, emotion/engagement, and voice perception in English and Mandarin readers, focusing on what the emerging “AI reading signature” suggests about cross-linguistic psycholinguistic processing and AI-mediated literary experience.

Across both language groups, AI-mediated stories produced a reliable reduction in narrative predictability as indexed by cloze performance, with the deficit substantially larger for English than Mandarin readers. Because cloze probability is a classic and widely validated behavioral proxy for contextual expectancy in language comprehension [23], this pattern supports the view that AI text—despite careful length and readability matching—creates subtly different predictive environments for readers. One plausible interpretation is that AI narratives may provide weaker or less stable distributional cues for forward prediction at critical junctures, consistent with broader predictive-processing accounts of real-time comprehension [24,25]. The cross-linguistic asymmetry we observed is also compatible with the idea that prediction is sensitive to language-specific conventions and training-ecology differences in contemporary AI text, although this remains an inference that should be directly tested with genre- and register-balanced corpora across languages.

In the subjective domain, AI-mediated texts also elicited lower narrative engagement and emotional intensity in both cohorts, with no reliable Text Type × Language interaction. This suggests a cross-culturally robust experiential penalty in AI reading that is not simply a byproduct of surface-level readability or length differences, but instead reflects higher-order narrative qualities tied to immersion and affective resonance. Our results align with established models and measures of narrative engagement that emphasize mental model construction, attentional focus, and affective involvement as core components of literary absorption [26]. They are also consistent with the broader transportation tradition in narrative psychology, where reduced immersion predicts weaker emotional and persuasive effects [27].

The strongest and most consistent AI deficits emerged for narrative voice perception. Human-authored texts were rated higher in authenticity/presence, stylistic naturalness, and perspectival coherence across both languages, with the largest standardized gaps centered on authenticity. This profile supports the idea that readers treat "voice" as a composite signal of human intentionality, experiential grounding, and stylistic idiosyncrasy that remains difficult for AI-mediated prose to fully reproduce under controlled matching conditions. Conceptually, this finding integrates well with research emphasizing that narrative quality and engagement are anchored not only in what a story conveys, but how an inferred authorial mind seems to inhabit the text [26,27]. The language-sensitive interaction we observed for coherence further suggests that maintaining a stable perspective may be a particularly salient cue for English readers or at least for the English narrative styles represented in our stimulus set again highlighting the need for future work that systematically varies viewpoint structures across languages and genres.

Our moderation analyses add an important nuance: AI familiarity attenuated subjective penalties in engagement, emotion, and voice (authenticity and naturalness), but did not significantly moderate the cloze-based predictability deficit. This dissociation implies that familiarity shapes evaluative and metacognitive judgments perhaps by recalibrating expectations of what AI writing can reasonably accomplish without substantially altering the lower-level predictive mechanics that readers deploy during online comprehension. This pattern resonates with emerging evidence that attitudes toward and disclosures of AI involvement can shift perceived quality and trust without necessarily changing the textual content itself [28,29]. In other words, readers may become more tolerant of AI voice or style with experience, yet still experience measurable predictability differences in the moment-to-moment processing of AI-mediated language.

Taken together, the effect-size synthesis and the NTDI framework suggest that AI-mediated narrative differences are multidimensional rather than unitary. The "signature" we observe is characterized by a comparatively large and cross-linguistically stable penalty in perceived voice authenticity and a consistent reduction in emotional engagement, coupled with a more language-sensitive disruption in predictive fluency. This triadic pattern offers a useful scaffold for future theorizing and for applied contexts such as AI-assisted publishing, translation, and educational storytelling, where the goal may be to improve not only grammaticality or coherence but the felt presence of an authorial voice.

5. CONCLUSION

This study investigated how readers process human-authored versus AI-mediated narratives across English and Mandarin, using a tightly matched stimulus set and a multi-domain framework spanning prediction, emotion/engagement, and narrative voice. With 12 matched story pairs and a large, balanced sample (N = 528; 264 per language), the design allowed us to isolate qualitative effects of AI mediation from basic surface-level confounds such as length and readability.

Across domains, a consistent "AI deficit" emerged. AI-mediated texts were less predictable in the cloze task, demonstrating a measurable disruption in forward narrative expectation. This effect was language-sensitive, with a larger predictability penalty for English readers than Mandarin readers. In contrast, emotional engagement and emotional intensity showed a strong main effect of Text Type without a cross-linguistic interaction, indicating that the felt experience of reading AI-mediated stories is reliably reduced across both cohorts. The most pronounced and theoretically central differences appeared in narrative voice, where AI texts were rated lower in authenticity/presence, stylistic naturalness, and perspectival coherence, with a language-specific amplification of the coherence deficit for English.

The integrative synthesis strengthens these conclusions. Effect-size comparisons indicated that voice-related penalties were generally the largest, while the cross-linguistic contrast was most visible in prediction. The Narrative Triad Divergence Index (NTDI) further captured this multidimensional pattern, showing a higher overall divergence for English than Mandarin readers, largely driven by differences in predictive disruption. Finally, the moderation analyses revealed a psychologically meaningful dissociation: AI familiarity attenuated subjective penalties (engagement, emotion, authenticity/naturalness) but did not reliably alter predictability, suggesting that experience with AI may recalibrate evaluation and expectations more than it reshapes lower-level processing dynamics.

Taken together, these findings suggest that AI-mediated literature is not experienced as merely "different style," but as a systematically distinct narrative signal that affects how readers anticipate, feel, and infer voice with some elements appearing cross-culturally stable and others dependent on language-specific reading norms or AI training ecologies. Practically, this implies that improving AI writing for literary contexts will require more than fluency: advances must target voice authenticity, stable perspective management, and predictive coherence, especially in English narrative settings. As AI-assisted storytelling becomes more common in publishing, education, and digital humanities, this triad framework offers a clear roadmap for evaluating progress and preserving the qualities readers most associate with human-mediated narrative experience.

REFERENCES

1. Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32–59. doi:10.1080/23273798.2015.1102299.
2. Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin, 144*(10), 1002–1044. doi:10.1037/bul0000158.
3. Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161–163. doi:10.1038/307161a0.
4. DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117–1121. doi:10.1038/nn1504.
5. Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302–319. doi: 10.1016/j.cognition.2013.02.013.
6. Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences.* doi:10.1073/pnas.2307876121.
7. Rayner, K. (2009). Eye movements in reading: Models and data. *Journal of Eye Movement Research, 2*(5), 1–10.
8. Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass, 9*(8), 311–327. doi:10.1111/lnc3.12151.
9. Frost, R. (2012). Towards a universal model of reading. *Behavioral and Brain Sciences, 35*(5), 263–329. doi:10.1017/S0140525X11001841.
10. Perfetti, C. A. (2003). The universal grammar of reading. *Scientific Studies of Reading, 7*(1), 3–24. doi:10.1207/S1532799XSSR0701_02.
11. Koda, K., & Zehler, A. M. (Eds.). (2008). *Learning to read across languages: Cross-linguistic relationships in first- and second-language literacy development.* Routledge.
12. Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology, 79*(5), 701–721. doi:10.1037/0022-3514.79.5.701.
13. Busselle, R., & Bilandzic, H. (2009). Measuring narrative engagement. *Media Psychology, 12*(4), 321–347. doi:10.1080/15213260903287259.
14. Mar, R. A., & Oatley, K. (2008). The function of fiction is the abstraction and simulation of social experience. *Perspectives on Psychological Science, 3*(3), 173–192. doi:10.1111/j.1745-6924.2008.00073. x.
15. Mar, R. A., Oatley, K., & Peterson, J. B. (2009). Exploring the link between reading fiction and empathy: Ruling out individual differences and examining outcomes. *Communications, 34*(4), 407–428. doi:10.1515/COMM.2009.025.
16. Bal, P. M., & Veltkamp, M. (2013). How does fiction read influence empathy? An experimental investigation on the role of emotional transportation. *PLOS ONE, 8*(1), e55341. doi: 10.1371/journal.pone.0055341.
17. Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science, 342*(6156), 377–380. doi:10.1126/science.1239918.
18. Genette, G. (1980). *Narrative discourse: An essay in method* (J. E. Lewin, Trans.). Cornell University Press.
19. Fludernik, M. (2009). *An introduction to narratology.* Routledge.
20. Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication, 25*(1), 89–100. doi:10.1093/jcmc/zmz022.
21. Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* doi:10.1145/3290605.3300469.
22. Berber Sardinha, T. (2024). AI-generated vs human-authored texts: A multidimensional comparison.

Applied Corpus Linguistics, 4(1), 100083. doi: 10.1016/j.acorp.2023.100083.

23. Taylor, W. L. (1953). *"Cloze Procedure": A New Tool for Measuring Readability.* Journalism Quarterly, 30(4), 415–433.
24. Kutas, M., & Hillyard, S. A. (1984). *Brain potentials during reading reflect word expectancy and semantic association.* Nature, 307, 161–163.
25. Gastaldon, S., et al. (2024). *Predictive language processing: integrating comprehension and production approaches.* Frontiers in Psychology.
26. Busselle, R., & Bilandzic, H. (2009). *Measuring Narrative Engagement.* Media Psychology, 12(4), 321–347.
27. Green, M. C., & Brock, T. C. (2000). *The Role of Transportation in the Persuasiveness of Public Narratives.* Journal of Personality and Social Psychology.
28. Lim, S., et al. (2024). *The effect of source disclosure on evaluation of AI- vs. human-generated messages.* (Journal article).
29. "How Text Presentation Influences Perceptions of AI Writing …" (2025). ACM conference paper.