Multi-Level Narrative Evaluation Outperforms Lexical Features for Mental Health

Yuxi Ma[1,2,3,4,5*], **Jieming Cui**[1,2,4,5*], **Muyang Li**[2*], **Ye Zhao**[2,6], **Yu Li**[2], **Yixuan Wang**[2], **Chi Zhang**[3,4✉], **Yinyin Zang**[2,5✉], **and Yixin Zhu**[2,1,4,5✉]

*Equal contributors Project Website: https://mayuxi.com/research/theraputic_writing
[1] Institute for Artificial Intelligence, Peking University [2] School of Psychological and Cognitive Sciences, Peking University
[3] School of Intelligence Science and Technology, Peking University [4] State Key Laboratory of General AI, Peking University
[5] Beijing Key Laboratory of Behavior and Mental Health, Peking University [6] PKU-Changsha Institute for Computing and Digital Economy

Abstract

How people narrate their experiences offers a window into how the mind organizes them. Computational approaches to therapeutic writing have evolved from lexical counting to neural methods, yet remain fragmented: dictionary tools miss discourse structure, while embeddings conflate local coherence with global organization. No existing framework maps these techniques onto the hierarchical processes through which narratives are constructed. Here we introduce a three-level framework—micro-level lexical features, meso-level semantic embeddings, and macro-level Large Language Model (LLM) narrative evaluation—and show, across 830 Chinese therapeutic texts spanning depression, anxiety, and trauma, that macro-level evaluation substantially outperforms lexical and embedding features for mental health prediction. This challenges the field's emphasis on word-counting: *formal structural features* (Labov's story grammar, Rhetorical Structure Theory (RST) coherence, propositional composition) demonstrate that narrative organization per se carries predictive signal, while *clinically-grounded narrative dimensions* capture how psychological states are expressed through discourse. Semantic embeddings add minimal independent value but yield incremental gains in multi-level classification. By grounding computational levels in discourse processing theory, this framework identifies macro-structural organization as the primary locus of clinical signal and generates testable hypotheses for intervention design and longitudinal research.

Keywords: computational linguistics; mental health; narrative coherence; semantic embeddings; large language models

Introduction

How we narrate our lives offers a window into how the mind organizes experience. Individuals experiencing depression, anxiety, or trauma construct stories that differ from healthy narratives not merely in content, but in structural organization. While coherent, integrated life narratives predict psychological well-being (Adler et al., 2016), therapeutic writing interventions demonstrate clinical efficacy across diverse samples (Pennebaker, 2016). Yet it remains unclear which levels of narrative construction carry the strongest mental health signal.

Current computational approaches are theoretically fragmented. Dictionary-based methods such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) capture word-level frequencies but overlook discourse structures central to therapeutic change (Boyd & Schwartz, 2021; Eid & Diener, 2006). Distributional embeddings (Mikolov et al., 2013) quantify semantic similarity yet conflate local coherence with global narrative organization. Deep neural classifiers achieve high accuracy but function as black boxes lacking interpretability (Teng et al., 2022). This fragmentation reflects a deeper theoretical absence: no framework maps these computational techniques onto the hierarchical processes through which narratives are constructed.

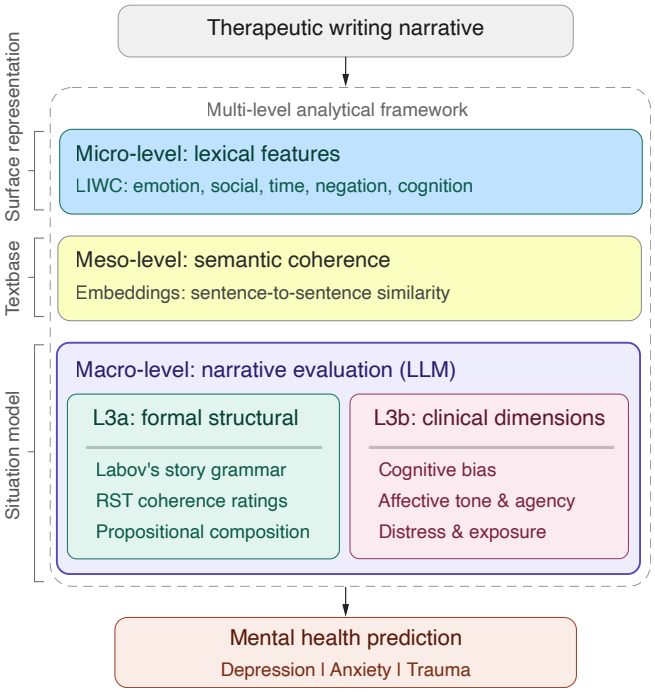


Figure 1: **Multi-level analytical framework for therapeutic writing.** Three computational layers—grounded in discourse processing theory (Kintsch & Van Dijk, 1978)—operationalize hierarchical narrative construction. The micro-level captures lexical patterns via LIWC; the meso-level quantifies semantic coherence through sentence embeddings; the macro-level employs LLMs as structured evaluators, distinguishing *formal structural features* (L3a: Labov's story grammar, RST coherence, propositional composition) from *clinically-grounded narrative dimensions* (L3b: cognitive bias, affective tone, distress). Left annotations map each layer onto established levels of text representation. All three layers feed into prediction of depression, anxiety, and trauma severity.

Established discourse processing models provide a natural scaffold. Kintsch and Van Dijk (1978) and subsequent work (Graesser et al., 2004; Van Dijk, Kintsch, et al., 1983) propose that text comprehension and production operate across three levels: surface representations (verbatim linguistic form), textbases (propositional meaning and local coherence), and situation models (global mental simulations). Mental health conditions are thought to affect narratives across all three levels—depression is associated with ruminative, self-focused surface language (Rude et al., 2004), trauma with fragmented

propositional flow (Brewin et al., 2010), and both with impoverished global narrative structure (Adler et al., 2016). Yet no existing approach operationalizes this hierarchy to analyze therapeutic writing.

We address this gap by integrating symbolic, distributional, and generative methods into a unified framework grounded in this three-level architecture (see Fig. 1). The micro-level layer employs Chinese LIWC to capture automatic lexical choices reflecting psycho-affective states. The meso-level layer uses text embeddings to quantify semantic coherence through sentence-to-sentence propositional flow. At the macro-level, LLMs function as structured evaluators extracting multi-dimensional narrative assessments. Critically, we distinguish two analytically separable sub-components within this layer: (i) *formal structural features*—Labov's story grammar components, RST coherence ratings, and compositional ratios—that capture the organizational architecture of the narrative; and (ii) *clinically-grounded narrative dimensions* that assess how psychological states (cognitive bias, distress, affective tone) are organized and expressed through discourse.

Analyzing 830 Chinese therapeutic writing samples from an ecologically diverse sample spanning ages 9–50 across clinical and community settings, we find that macro-level evaluation substantially outperforms lexical and semantic features for predicting depression, anxiety, and trauma severity. We make three contributions: (i) a theoretically grounded multi-level framework that aligns computational methods with established discourse processing theory; (ii) evidence that narrative organization, not vocabulary, is the primary carrier of mental health signal in therapeutic writing; and (iii) condition-specific narrative signatures (*e.g.*, temporal disorganization in depression, spatial grounding deficits in anxiety) that generate testable hypotheses for future intervention and longitudinal research.

Related Work

Therapeutic writing and mental health Expressive writing about emotional experiences produces measurable improvements in psychological and physical health (Pennebaker, 2016), with meta-analyses demonstrating moderate effect sizes across depression, anxiety, and trauma-related disorders (Frattaroli, 2006). The dominant tool for quantifying these shifts, LIWC (Pennebaker et al., 2015), categorizes words into psycholinguistic dimensions and has identified reliable linguistic markers—elevated negative emotion words and first-person singular pronouns in depression, increased causal and cognitive processing terms during recovery (Rude et al., 2004). However, LIWC's reliance on fixed word lists prevents it from capturing context-dependent meanings, and its bag-of-words architecture neglects the semantic relationships and discourse structures that are central to therapeutic change (Taraban & Abusal, 2019). These limitations motivate approaches that move beyond surface-level word counting.

Semantic coherence and computational analysis One such direction targets semantic coherence, the local integration of adjacent discourse units, serving as a clinical marker of cognitive-linguistic stability. Disorganized propositional flow and ruminative fragmentation are characteristic of both depression and trauma-related symptoms (Brewin et al., 2010; Nolen-Hoeksema, 1991). Computational methods have quantified coherence through lexical overlap (Halliday & Hasan, 1976), entity grids (Barzilay & Lapata, 2008), and semantic vector similarity (Zirikly et al., 2019), but these features predominantly capture local, sentence-to-sentence transitions. A persistent gap remains: existing approaches often conflate such meso-level transitions with macro-level narrative organization, failing to distinguish the maintenance of local semantic flow from the global construction of narrative meaning.

LLMs as structured evaluators in mental health LLMs offer a potential path beyond this conflation. Their application in computational psychiatry has expanded rapidly, spanning suicide risk assessment, psychological consulting, and clinical text analysis (Sharma et al., 2023; Song et al., 2025; Yang et al., 2023). Yet most LLM applications remain end-to-end classifiers that inherit the "black-box" opacity of deep learning (Guo et al., 2024). An emerging alternative employs LLMs not as classifiers but as structured evaluators, using prompt engineering to operationalize theoretically grounded dimensions (Guo et al., 2024; Stade et al., 2024). This zero-shot evaluative paradigm has shown promise in educational assessment (Mizumoto & Eguchi, 2023) and cognitive bias detection (Ke et al., 2024), but its potential for decoding the macro-structural organization of clinical narratives—where global coherence and rhetorical logic are paramount—remains largely unexplored.

Methods

Dataset and Participants

We analyzed 830 Chinese therapeutic writing samples (> 100 words) from individuals aged 9 to 50 years ($M = 20.7$, $SD = 7.7$; 76.4% female) who completed 20–30 minute expressive writing sessions about emotionally significant experiences. Samples were drawn from six therapeutic interventions conducted between 2018 and 2024, spanning clinical, post-disaster, school-based, and online settings. The clinical adult sample comprised Writing Exposure Therapy (M. Li, Zhao, Guo, et al., 2025; $n = 30$; $M_{\text{age}} = 28.6 \pm 7.9$; 73.3% female) and Guided Writing Exposure (M. Li, Zhao, Rosenfield, et al., 2025; $n = 84$; $M_{\text{age}} = 27.0 \pm 6.0$; 86.9% female). Post-disaster child and adolescent samples included the Sichuan Earthquake Children study ($n = 159$; $M_{\text{age}} = 11.1 \pm 1.2$; 52.8% female) and the Jishishan Group Intervention ($n = 43$; $M_{\text{age}} = 14.3 \pm 0.8$; 62.8% female). Additional samples came from Hubei Primary Students preventive interventions ($n = 50$; $M_{\text{age}} = 10.0 \pm 0.2$; 74.0% female) and Tencent Medical Platform online users ($n = 464$; $M_{\text{age}} = 24.0 \pm 5.1$; 84.3% female). Across the pool, 137 participants (16.5%) met criteria for clinical depression, 100 (12.0%) for clinical anxiety, and 323 (38.9%) for probable Post-Traumatic Stress Disorder (PTSD).

Depression was measured using the Beck Depression Inventory-II (BDI-II; Beck et al., 1996), Patient Health

Questionnaire-9 (PHQ-9; Kroenke et al., 2001), Patient Health Questionnaire-4 (PHQ-4; Kroenke et al., 2009), or the Revised Child Anxiety and Depression Scale (RCADS-47 or RCADS-25; Chorpita et al., 2000; Ebesutani et al., 2012). Anxiety was assessed via the Beck Anxiety Inventory (BAI; Beck et al., 1988), Generalized Anxiety Disorder-7 (GAD-7; Spitzer et al., 2006), PHQ-4, or RCADS. Trauma symptoms were evaluated using the PTSD Symptom Scale Interview for DSM-5 (PSSI-5; Foa et al., 2016), Child PTSD Symptom Scale for DSM-5 – Interview Version (CPSS-5-I; Foa et al., 2018), Children's Revised Impact of Event Scale (CRIES; Smith et al., 2003), or Primary Care PTSD Screen (PC-PTSD-5; Prins et al., 2016). Because the six sub-studies employed different instruments, raw scores were normalized to a 0–1 range by dividing by each instrument's maximum possible score. This linear rescaling preserves within-instrument rank ordering while enabling cross-study pooling, though it assumes approximate comparability of severity thresholds across instruments—a simplification we acknowledge as a limitation. Normalized scores were then converted into ordinal severity levels: four for depression and anxiety (none, mild, moderate, severe) and five for trauma (none, mild, moderate, severe, very severe). All participants provided informed consent; all contributing studies received institutional ethics approval.

Multi-Level Analytical Framework

Layer 1: Micro-level (lexical features) The micro-level captures linguistic patterns through theory-driven feature selection from Simplified Chinese LIWC (Gao et al., 2013). Each feature maps to a specific psychological mechanism: *first-person singular pronouns* reflect self-focused attention (Pyszczynski & Greenberg, 1987); *negative emotion words* index negative cognitive bias (Beck et al., 1979); certitude and discrepancy markers (*certain*, *discrep*) capture absolutist thinking versus recognition of situational gaps (Egan et al., 2011); *social words* indicate interpersonal engagement (Teo et al., 2013); *past-tense focus* marks rumination (Nolen-Hoeksema, 1991); *death-related terms* signal suicidal ideation (Pennebaker et al., 2015); and *negation words* (*negate*) proxy defensive cognitive operations (Pennebaker, 1997).

Layer 2: Meso-level (semantic coherence) The meso-level quantifies semantic integration using OpenAI's `text-embedding-3-small` model, generating 1536-dimensional vectors for each sentence and the complete text. *Local coherence* was operationalized via sentence-to-sentence (s2s) cosine similarity, yielding indices of average propositional flow (s2s_mean), abrupt thematic shifts (s2s_min), and transition variability (s2s_std). *Global coherence* was captured through sentence-to-document (s2d) similarity, measuring thematic alignment (s2d_mean), wandering (s2d_std), and representative extremes (s2d_max/min). Local measures reflect the fluidity of adjacent idea transitions; global measures index the narrator's ability to anchor propositions to a central theme (J. Li & Hovy, 2014).

Layer 3: Macro-level (narrative evaluation) At the macro-level, GPT-4o functions as a structured evaluator, extracting multi-dimensional narrative assessments via deterministic sampling (temperature = 0) with JSON-formatted outputs and supporting textual evidence, ensuring reproducibility and external auditability. We distinguish two sub-components within this layer, operationalized through three complementary evaluation protocols.

L3a: Formal structural features This sub-component captures the organizational architecture of the narrative independent of clinical content, via two evaluation protocols: (i) *Propositional and rhetorical logic.* Guided by macrostructure theory (Van Dijk, 2019), LLMs decomposed narratives into minimal semantic units—each comprising a core predicate and its subject—and categorized them into Actions & Facts, Sensory Perception, Direct Emotion, Indirect Emotion, and Cognition. These were aggregated into three functional dimensions: cognitive processing, affective engagement, and narrative grounding, capturing the allocation of conceptual resources during writing (Bruner, 1991). Concurrently, following RST (Mann & Thompson, 1988), LLMs identified rhetorical relations (*e.g.*, Elaboration, Cause, Contrast) and assessed transition quality on 0–5 scales, aggregated into a global coherence score. (ii) *Canonical narrative organization.* LLMs evaluated texts against Labov's story grammar (Labov, 1972), rating 6 structural components (abstract, orientation, complicating action, evaluation, resolution, and coda) on 5-point scales with supporting evidence. These scores index the narrator's capacity for coherent macrostructure and meaning-making, both robust predictors of psychological well-being (Adler, 2012; Adler et al., 2016).

L3b: Clinically-grounded narrative dimensions This sub-component assesses how mental health states are expressed and organized via narrative discourse. LLMs evaluated 15 dimensions drawn from trauma-focused CBT (Cohen et al., 2006) and cognitive therapy models (Beck et al., 1979), organized into 4 categories: (i) *structural trauma processing* examines whether the narrative macrostructure supports exposure acceptance or manifests avoidance, overgeneralization, and episodic specificity loss (Williams et al., 2007); (ii) *cognitive processing* identifies how structural organization expresses cognitive biases and depth of sense-making; (iii) *affective/agentic integration* measures agency and affective tone within the story arc; and (iv) *global structural coherence* assesses spatio-temporal consistency and contextual density. Crucially, these dimensions capture the *narrative expression* of psychological states—how cognitive patterns and affective experiences are organized within discourse—rather than direct symptom reports. For instance, `cognitive_bias_score` reflects how reasoning patterns structure narrative (*e.g.*, the degree to which causal attributions, absolutist language, and negative schema are narratively organized), not a questionnaire response about thought frequency.

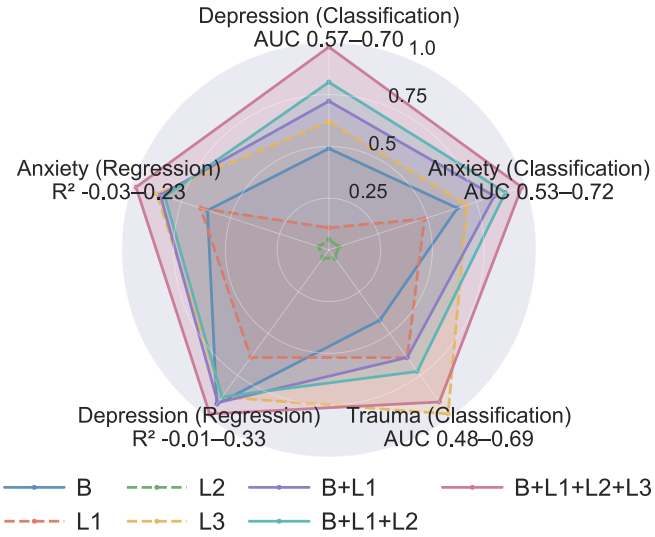


Figure 2: **Performance across layered feature sets.** Radar plot comparing regression (R^2) and classification (AUC) performance across five tasks. Each line represents a feature combination, from the baseline alone (B) through sequential layer addition up to $B + L1 + L2 + L3$. All metrics are normalized to a maximum of 1.00. Layer 3 (macro-level) alone approaches full-model performance, while Layer 2 (meso-level) alone falls near the plot center, indicating near-chance accuracy. Sequential integration shows incremental gains for classification but plateauing R^2 until Layer 3 is added.

Evaluation and Metrics

We evaluated predictive utility through two complementary tasks: continuous symptom regression (depression and anxiety) and multi-class severity classification (depression, anxiety, and trauma). Trauma was excluded from regression due to the absence of continuous symptom scores in the source datasets. Regression employed *ExtraTreesRegressor*; classification used *Gradient Boosting* and *ExtraTrees* with class-balanced weights to accommodate severity-level imbalance. All models were validated via stratified 5-fold cross-validation across 7 feature combinations—from baseline demographics (age and gender) alone to the full B+L1+L2+L3 suite—enabling systematic assessment of each layer's marginal contribution. Given minimal L3 missingness (0.4%), zero-filling preserved cross-model comparability without introducing synthetic bias. Regression performance was assessed with R^2, Root Mean Square Error (RMSE), and Mean Absolute Error (MAE); classification with Area Under the Curve (AUC), Balanced Accuracy, and Macro-F1.

Results

Model Performance Across Feature Combinations

Performance varied substantially across computational layers and their combinations (Tab. 1 and Fig. 2). The full model (B+L1+L2+L3) achieved the best performance across all tasks: depression $R^2 = 0.332$ and anxiety $R^2 = 0.235$ for regression; depression AUC=0.699, anxiety AUC=0.718, and trauma AUC=0.676 for classification.

When evaluated in isolation, the three layers exhibited a clear hierarchy. Layer 3 (macro-level) was the strongest single predictor, approaching full-model performance (depression $R^2 = 0.295$, anxiety $R^2 = 0.204$, trauma AUC=0.692) and substantially outperforming baseline demographics. Layer 1 (micro-level) showed moderate standalone capability comparable to the baseline. Layer 2 (meso-level) demonstrated negligible independent utility, yielding negative regression scores ($R^2 = -0.014$ for depression, -0.034 for anxiety) and near-chance classification (AUC$\leq$0.566 across all conditions)—indicating that isolated semantic coherence features fail to capture meaningful variance beyond a mean-only baseline.

Sequential layer integration, however, revealed a more nuanced picture. Adding L1 to baseline consistently improved classification (depression AUC: 0.629→0.662; anxiety: 0.651→0.690; trauma: 0.567→0.617), and incorporating L2 yielded further incremental gains (depression: 0.675; anxiety: 0.701; trauma: 0.635)—despite L2's poor standalone performance. This suggests that semantic coherence features capture complementary variance when combined with lexical and demographic information. The two task types showed divergent integration trajectories: classification accuracy improved incrementally with each added layer, whereas regression R^2 plateaued across B, B+L1, and B+L1+L2 combinations, increasing substantially only upon inclusion of Layer 3.

Feature Importance Patterns

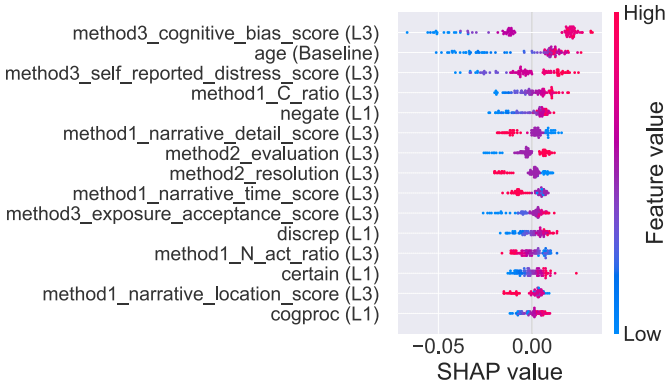
SHapley Additive exPlanations (SHAP) analysis revealed hierarchical importance patterns across the three computational layers (Fig. 3). The top 15 features for depression and anxiety showed substantial overlap (11 shared predictors), spanning L3 narrative assessments, L1 lexical markers, and baseline demographics.

Clinically-grounded narrative dimensions (L3b) ranked highest within L3. `cognitive_bias_score` emerged as the single strongest predictor for both depression (SHAP = 0.024) and anxiety (SHAP = 0.019), followed by `self_reported_distress_score` (3rd for both). These features capture how cognitive distortions and affective distress are organized and elaborated in narrative discourse, rather than direct responses to symptom questionnaires.

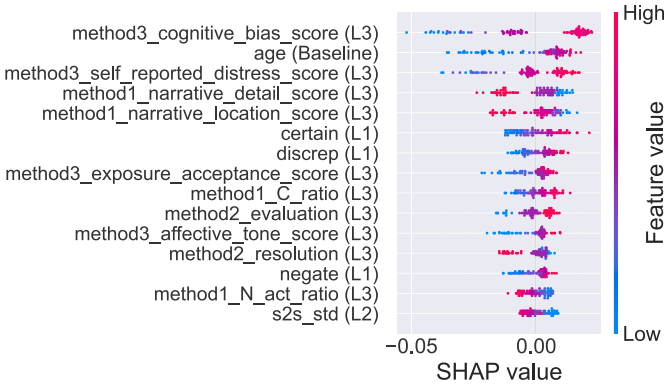
Formal structural features (L3a) provided convergent ev-

Table 1: **Performance comparison across feature combinations.** R^2 is reported for regression (depression and anxiety); AUC for multi-class classification (depression, anxiety, and trauma). Values denote mean±SD across stratified 5-fold cross-validation. Individual layers are shown above the rule; cumulative combinations below. Bold indicates best performance per column. Trauma regression is omitted due to the absence of continuous scores in source datasets.

Feature Set	**Regression (R^2)↑**		**Classification (AUC)↑**		
	Dep.	Anx.	Dep.	Anx.	Trauma
Baseline (B)	$0.311^{\pm 0.115}$	$0.130^{\pm 0.050}$	$0.629^{\pm 0.017}$	$0.651^{\pm 0.047}$	$0.567^{\pm 0.040}$
L1	$0.207^{\pm 0.030}$	$0.140^{\pm 0.048}$	$0.574^{\pm 0.014}$	$0.617^{\pm 0.037}$	$0.617^{\pm 0.075}$
L2	$-0.014^{\pm 0.039}$	$-0.034^{\pm 0.019}$	$0.566^{\pm 0.058}$	$0.527^{\pm 0.058}$	$0.484^{\pm 0.076}$
L3	$0.295^{\pm 0.021}$	$0.204^{\pm 0.045}$	$0.647^{\pm 0.017}$	$0.661^{\pm 0.039}$	$0.692^{\pm 0.062}$
B + L1	$0.308^{\pm 0.027}$	$0.199^{\pm 0.037}$	$0.662^{\pm 0.014}$	$0.690^{\pm 0.022}$	$0.617^{\pm 0.064}$
B + L1 + L2	$0.294^{\pm 0.036}$	$0.192^{\pm 0.034}$	$0.675^{\pm 0.021}$	$0.701^{\pm 0.021}$	$0.635^{\pm 0.079}$
B + L1 + L2 + L3	$\mathbf{0.332^{\pm 0.020}}$	$\mathbf{0.235^{\pm 0.039}}$	$\mathbf{0.699^{\pm 0.018}}$	$\mathbf{0.718^{\pm 0.040}}$	$\mathbf{0.676^{\pm 0.085}}$



(a) **Depression.** `cognitive_bias_score` (L3b) and age (Baseline) are the two most influential predictors. Higher cognitive bias scores drive increased predicted severity; higher Labovian `evaluation` and `resolution` scores (L3a) are associated with lower severity.



(b) **Anxiety.** `cognitive_bias_score` (L3b) again ranks first, followed by age and `self_reported_distress_score` (L3b). Anxiety-specific predictors include spatial grounding (`narrative_location_score`, L3a) and uncertainty markers (`certain`, `discrep`, L1).

Figure 3: **SHAP summary plots for depression and anxiety score prediction.** Features are ranked top-to-bottom by mean absolute SHAP value (global importance). Each dot represents one sample; horizontal position indicates the SHAP value (contribution to model output), and color encodes the original feature magnitude (red = high, blue = low). Feature labels indicate their computational layer: L3a (formal structural), L3b (clinically-grounded narrative dimensions), L1 (lexical), L2 (semantic coherence), or Baseline (demographics). Of the top 15 predictors, 11 are shared across both conditions, with L3 features dominating.

idence that narrative organization carries predictive signal independent of expressed clinical content. Labovian components (`evaluation`, `resolution`) and propositional organization scores (`narrative_time_score`, `C_ratio`, `narrative_detail_score`) appeared consistently in the top 15 for both conditions, with a negative association with symptom severity: narratives exhibiting greater meaning-making capacity, temporal organization, and compositional balance predicted lower distress, consistent with clinical theories of narrative integration (Adler, 2012; Adler et al., 2016).

Age ranked second overall (depression SHAP = 0.017; anxiety SHAP = 0.012). Given that the six sub-samples span ages 9–50 and differ systematically in therapeutic context, age in the pooled model likely partially indexes population membership and baseline severity rather than developmental effects alone—an interpretive constraint we address in the Discussion.

Despite this shared core, condition-specific signatures emerged. Depression models favored cognitive composition (`C_ratio`) and temporal organization (`narrative_time_score`), while anxiety models prioritized uncertainty markers (`certain`, `discrep`) and spatial grounding (`narrative_location_score`). Across both conditions, protective narrative features—structural coherence, compositional balance, narrative detail, and Labovian closure—consistently predicted lower symptom scores. Layer 2 semantic coherence features showed minimal predictive utility: all metrics fell outside the top 15 for depression, and only `s2s_std` appeared at rank 15 for anxiety (SHAP = 0.004), suggesting that variability in sentence-to-sentence semantic flow may specifically characterize anxious narratives.

Discussion

This study introduced a theoretically grounded, multi-level computational framework that integrates symbolic, distributional, and generative approaches for analyzing mental health states in therapeutic writing. Across 830 Chinese narratives spanning depression, anxiety, and trauma, three key findings emerged: macro-level narrative features were the dominant predictor of symptom severity, semantic coherence contributed minimally in isolation but added value in combined models, and sequential layer integration revealed task-dependent patterns—classification benefited incrementally from each layer, while regression improved substantially only upon inclusion of macro-level structural features.

Macro-Level Narrative Evaluation Outperforms Lexical and Embedding Approaches

Our central finding—that macro-level narrative (L3) features explain substantially more variance than lexical or semantic features—challenges the historical emphasis on word-counting in computational psycholinguistics. This suggests that mental health status is not merely associated with the frequency of specific distress words, but is more strongly linked to the hierarchical organization of the narrator's situation model (Kintsch & Van Dijk, 1978). The performance superiority of Layer 3 indicates that **the way** a story is constructed is more informative than the specific vocabulary or textbase representations (L2 embeddings) used to tell it.

Within the macro-level layer, formal structural features (L3a) provide convergent evidence that narrative organization carries predictive signal independent of expressed content. Elements such as `narrative_time_score` and `narrative_location_score` capture the narrator's capacity to maintain a stable spatiotemporal field. Our results are consistent with the possibility that psychological distress

is associated with a breakdown in this scaffolding, where high cognitive load or trauma-related dissociation may impair the construction of a coherent situational model. This pattern is particularly salient in high-severity samples, where narrative structural integrity tends to give way to disjointed propositions. These formal properties index the narrator's capacity for meaning-making closure, a robust predictor of psychological well-being (Adler, 2012) that remains invisible to traditional lexical or embedding-based metrics.

Concurrently, the predictive power of clinically-grounded narrative dimensions (L3b) suggests that psychopathology may function as a structural filter that reshapes narrative output. Rather than appearing as isolated errors, cognitive biases (*e.g.*, catastrophizing, all-or-nothing thinking, and overgeneralization) appear to operate as systemic organizing principles that shape the narrative arc and causal reasoning. These dimensions may reflect the expansion of negative self-schema (Beck et al., 1979), wherein idiosyncratic failures are codified into universal laws within the text. This suggests that LLMs are not merely detecting distress markers in the text, but are capturing how distress is organized and expressed across the discourse.

Importantly, the condition-specific signatures identified through SHAP analysis offer theoretically interpretable contrasts. Depression models favored cognitive composition (`C_ratio`) and temporal organization (`narrative_time_score`), consistent with the ruminative temporal disorientation characteristic of depressive cognition (Nolen-Hoeksema, 1991). Anxiety models, by contrast, prioritized uncertainty markers (*certain*, *discrep*) and spatial grounding (`narrative_location_score`), aligning with the hypervigilance and environmental scanning associated with anxiety disorders. These differential patterns generate testable hypotheses: if temporal scaffolding is specifically disrupted in depression, interventions that explicitly structure narrative chronology may prove more effective for depressive symptoms than generic expressive writing.

Methodological Contributions and the Challenge of Semantic Coherence

This work advances computational psycholinguistics through three methodological contributions. First, we replace opaque "black-box" extractions with a theory-driven framework where each feature maps to specific psychological hypotheses. By operationalizing micro-level (LIWC), semantic (embeddings), and macro-structural (RST, Labovian, Cognitive Behavioral Therapy (CBT)) dimensions, this multi-level approach provides a transparent and interpretable link between linguistic form and psychological state.

Second, employing LLMs as structured evaluators—rather than end-to-end classifiers—enhances analytical transparency. By requiring JSON-formatted outputs with supporting textual evidence, this framework ensures that Layer 3 assessments are externally auditable. This approach provides a rigorous alternative to open-ended generative assessment, bridging the gap between qualitative clinical judgment and quantitative prediction.

Third, the multi-level integration reveals a hierarchical dominance in clinical signaling: macro-structural features capture the core variance associated with psychological distress, whereas surface-level markers provide only peripheral information. This decomposition identifies the boundaries of each linguistic level, demonstrating that structural organization is a more robust correlate of psychopathology than lexical or semantic density in clinical narratives.

Notably, Layer 2's limited predictive power despite its theoretical prominence warrants reflection. While embedding-based similarity captures semantic overlap, it fails to model the sequential logic of psychological coherence. We hypothesize that in therapeutic writing, task constraints force narratives to cluster around trauma-related concepts, creating a ceiling effect for Layer 2 variance: all participants write about distressing experiences, compressing the semantic space in which embeddings operate. Layer 3, by contrast, reveals dramatic differences in *how* these shared themes are structurally organized. This interpretation suggests that distributional metrics may retain diagnostic value in unconstrained corpora (*e.g.*, social media), where topic variance is higher, but that structured clinical narratives require more granular operationalizations—such as discourse relation classifiers, argument structure analysis, or concept network graphs—to capture meaningful disruptions at the meso-level.

Limitations and Future Directions

Several limitations warrant consideration. First, sample heterogeneity across ages (9–50), therapeutic contexts, and assessment instruments introduces variance that age-as-covariate only partially addresses; the prominence of age in SHAP rankings likely reflects sub-sample differences rather than pure developmental effects. Second, while we distinguish L3a and L3b conceptually, the current analysis reports them as a combined set; future ablation studies—coupled with validation against human expert ratings—are needed to clarify if these LLM-generated signals reflect genuine psychological constructs or redundant surface features like text length. Third, our cross-sectional design precludes causal inference: the associations between narrative disorganization and higher severity are equally consistent with distress causing fragmentation, fragmented capacity maintaining distress, or both reflecting a shared process. Longitudinal studies tracking within-person narrative evolution across treatment could illuminate causal direction. Fourth, narrative conventions and coherence markers are culturally contingent, and whether patterns in our Chinese sample generalize to typologically different languages remains an open question. Finally, translational impact requires experimental validation—the condition-specific signatures we identified suggest that scaffolding particular narrative dimensions (temporal organization for depression, spatial grounding for anxiety) could enhance therapeutic writing, a hypothesis testable through randomized controlled trials.

Acknowledgement Y. Ma, J. Cui, and Y. Zhu are supported by the National Natural Science Foundation of China (32595491, 62376009), the PKU-BingJi Joint Laboratory for Artificial Intelligence, the Wuhan Major Scientific and Technological Special Program (2025060902020304), the Hubei Embodied Intelligence Foundation Model Research and Development Program, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone. M. Li, Y. Zhao, Y. Li, Y. Wang, and Y. Zang are supported by the National Natural Science Foundation of China (32371139, 32000776), and the Open Funding of the National Key Laboratory of Cognitive Neuroscience and Learning (CNLZD2103).

References

Adler, J. M. (2012). Living into the story: Agency and coherence in a longitudinal study of narrative identity development and mental health over the course of psychotherapy. *Journal of Personality and Social Psychology*, *102*(2), 367 (cit. on pp. 3, 5, 6).

Adler, J. M., Lodi-Smith, J., Philippe, F. L., & Houle, I. (2016). The incremental validity of narrative identity in predicting well-being: A review of the field and recommendations for the future. *Personality and Social Psychology Review*, *20*(2), 142–175 (cit. on pp. 1–3, 5).

Barzilay, R., & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics*, *34*(1), 1–34 (cit. on p. 2).

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, *56*(6), 893 (cit. on p. 3).

Beck, A. T., Rush, A. J., Shaw, B. F., Emery, G., DeRubeis, R. J., & Hollon, S. D. (1979). *Cognitive therapy of depression*. Guilford Press. (Cit. on pp. 3, 6).

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of beck depression inventories-ia and-ii in psychiatric outpatients. *Journal of Personality Assessment*, *67*(3), 588–597 (cit. on p. 2).

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21–41 (cit. on p. 1).

Brewin, C. R., Gregory, J. D., Lipton, M., & Burgess, N. (2010). Intrusive images in psychological disorders: Characteristics, neural mechanisms, and treatment implications. *Psychological Review*, *117*(1), 210 (cit. on p. 2).

Bruner, J. (1991). The narrative construction of reality. *Critical inquiry*, *18*(1), 1–21 (cit. on p. 3).

Chorpita, B. F., Yim, L., Moffitt, C., Umemoto, L. A., & Francis, S. E. (2000). Assessment of symptoms of dsm-iv anxiety and depression in children: A revised child anxiety and depression scale. *Behaviour Research and Therapy*, *38*(8), 835–855 (cit. on p. 3).

Cohen, J., Mannarino, A., Deblinger, E., et al. (2006). Treating trauma and traumatic grief in children and adolescents. *Guilford Publications* (cit. on p. 3).

Ebesutani, C., Reise, S. P., Chorpita, B. F., Ale, C., Regan, J., Young, J., Higa-McMillan, C., & Weisz, J. R. (2012). The revised child anxiety and depression scale-short version: Scale reduction via exploratory bifactor modeling of the broad anxiety factor. *Psychological Assessment*, *24*(4), 833 (cit. on p. 3).

Egan, S. J., Wade, T. D., & Shafran, R. (2011). Perfectionism as a transdiagnostic process: A clinical review. *Clinical Psychology Review*, *31*(2), 203–212 (cit. on p. 3).

Eid, M. E., & Diener, E. E. (2006). *Handbook of multimethod measurement in psychology*. American Psychological Association. (Cit. on p. 1).

Foa, E. B., Asnaani, A., Zang, Y., Capaldi, S., & Yeh, R. (2018). Psychometrics of the child ptsd symptom scale for dsm-5 for trauma-exposed children and adolescents. *Journal of Clinical Child & Adolescent Psychology*, *47*(1), 38–46 (cit. on p. 3).

Foa, E. B., McLean, C. P., Zang, Y., Zhong, J., Rauch, S., Porter, K., Knowles, K., Powers, M. B., & Kauffman, B. Y. (2016). Psychometric properties of the posttraumatic stress disorder symptom scale interview for dsm–5 (pssi–5). *Psychological Assessment*, *28*(10), 1159 (cit. on p. 3).

Frattaroli, J. (2006). Experimental disclosure and its moderators: A meta-analysis. *Psychological Bulletin*, *132*(6), 823 (cit. on p. 2).

Gao, R., Hao, B., Li, H., Gao, Y., & Zhu, T. (2013). Developing simplified chinese psychological linguistic analysis dictionary for microblog. *International Conference on Brain and Health Informatics* (cit. on p. 3).

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202 (cit. on p. 1).

Guo, Z., Lai, A., Thygesen, J. H., Farrington, J., Keen, T., Li, K., et al. (2024). Large language models for mental health applications: Systematic review. *JMIR Mental Health*, *11*(1), e57400 (cit. on p. 2).

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in english*. Routledge. (Cit. on p. 2).

Ke, Y., Yang, R., Lie, S. A., Lim, T. X. Y., Ning, Y., Li, I., Abdullah, H. R., Ting, D. S. W., & Liu, N. (2024). Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: Simulation study. *Journal of Medical Internet Research*, *26*, e59439 (cit. on p. 2).

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363 (cit. on pp. 1, 5).

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The phq-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613 (cit. on p. 3).

Kroenke, K., Spitzer, R. L., Williams, J. B., & Löwe, B. (2009). An ultra-brief screening scale for anxiety and depression: The phq–4. *Psychosomatics*, *50*(6), 613–621 (cit. on p. 3).

Labov, W. (1972). *Language in the inner city: Studies in the black english vernacular* (Vol. 3). University of Pennsylvania Press. (Cit. on p. 3).

Li, J., & Hovy, E. (2014). A model of coherence based on distributed sentence representation. *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)* (cit. on p. 3).

Li, M., Zhao, Y., Guo, Z., Wei, M., Fan, S., Chen, Q., Li, Y., & Zang, Y. (2025). Written exposure therapy for posttraumatic stress disorder and integration of a mindfulness based app in china: A pilot randomized controlled trial. *Behavior Therapy* (cit. on p. 2).

Li, M., Zhao, Y., Rosenfield, D., Guo, Z., Wei, M., Fan, S., Li, Y., & Zang, Y. (2025). An online guided written exposure therapy for symptoms of posttraumatic stress disorder: A randomized controlled trial. *Psychotherapy and Psychosomatics* (cit. on p. 2).

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, *8*(3), 243–281 (cit. on p. 3).

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* (cit. on p. 1).

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050 (cit. on p. 2).

Nolen-Hoeksema, S. (1991). Responses to depression and their effects on the duration of depressive episodes. *Journal of Abnormal Psychology*, *100*(4), 569 (cit. on pp. 2, 3, 6).

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, *8*(3), 162–166 (cit. on p. 3).

Pennebaker, J. W. (2016). *Opening up by writing it down: How expressive writing improves health and eases emotional pain*. Guilford Publications. (Cit. on pp. 1, 2).

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of liwc2015 (cit. on pp. 1–3).

Prins, A., Bovin, M. J., Smolenski, D. J., Marx, B. P., Kimerling, R., Jenkins-Guarnieri, M. A., Kaloupek, D. G., Schnurr, P. P., Kaiser, A. P., Leyva, Y. E., et al. (2016). The primary care ptsd screen for dsm-5 (pc-ptsd-5): Development and evaluation within a veteran primary care sample. *Journal of General Internal Medicine*, *31*(10), 1206–1211 (cit. on p. 3).

Pyszczynski, T., & Greenberg, J. (1987). Self-regulatory perseveration and the depressive self-focusing style: A self-awareness theory of reactive depression. *Psychological Bulletin*, *102*(1), 122 (cit. on p. 3).

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion*, *18*(8), 1121–1133 (cit. on pp. 1, 2).

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, *5*(1), 46–57 (cit. on p. 2).

Smith, P., Perrin, S., Dyregrov, A., & Yule, W. (2003). Principal components analysis of the impact of event scale with children in war. *Personality and Individual Differences*, *34*(2), 315–322 (cit. on p. 3).

Song, I., Pendse, S. R., Kumar, N., & De Choudhury, M. (2025). The typing cure: Experiences with large language model chatbots for mental health support. *ACM Conference on Human Factors in Computing Systems (CHI)* (cit. on p. 2).

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The gad-7. *Archives of Internal Medicine*, *166*(10), 1092–1097 (cit. on p. 3).

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. *NPJ Mental Health Research*, *3*(1), 12 (cit. on p. 2).

Taraban, R., & Abusal, K. (2019). Analyzing topic differences, writing quality, and rhetorical context in college students' essays using linguistic inquiry and word count (liwc). *East European Journal of Psycholinguistics* (cit. on p. 2).

Teng, Q., Liu, Z., Song, Y., Han, K., & Lu, Y. (2022). A survey on the interpretability of deep learning in medical diagnosis. *Multimedia Systems*, *28*(6), 2335–2355 (cit. on p. 1).

Teo, A. R., Choi, H., & Valenstein, M. (2013). Social relationships and depression: Ten-year follow-up from a nationally representative study. *PloS one*, *8*(4), e62396 (cit. on p. 3).

Van Dijk, T. A. (2019). *Macrostructures: An interdisciplinary study of global structures in discourse, interaction, and cognition*. Routledge. (Cit. on p. 3).

Van Dijk, T. A., Kintsch, W., et al. (1983). *Strategies of discourse comprehension*. Academic press New York. (Cit. on p. 1).

Williams, J. M. G., Barnhofer, T., Crane, C., Herman, D., Raes, F., Watkins, E., & Dalgleish, T. (2007). Autobiographical memory specificity and emotional disorder. *Psychological bulletin*, *133*(1), 122 (cit. on p. 3).

Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., & Ananiadou, S. (2023). Towards interpretable mental health analysis with large language models. *Annual Conference on Empirical Methods in Natural Language Processing (EMNLP)* (cit. on p. 2).

Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. *The Sixth Workshop on Computational Linguistics and Clinical Psychology* (cit. on p. 2).