The Clustering Paradox in Cross-Lingual Risk Expression: Distributional Universality and Temporal Necessity

Anonymous ACL submission

Abstract

Mental health chatbots increasingly serve users across languages, yet most risk-detection systems rely on English single-turn social media and lack temporal grounding. Multi-turn counseling data remain scarce due to privacy and ethical barriers. We address this gap using Korean professional counseling transcripts (N=2,833 dialogues) with turn-level risk annotations and English mental-health posts (N=3,512) for cross-lingual comparison. Using multilingual embeddings and optimal-transport distance, we find strong distributional universality between Korean and English risk expressions ($W < 0.02$) but absent categorical structure (ARI ≈ 0), indicating risk forms a continuous semantic spectrum. Supervised single-turn classification (F1=0.77) requires temporal aggregation for safety-critical deployment. Survival analysis of Korean dialogues establishes Minimum Safety Windows (MSW): $\text{MSW}_{0.5}$ = 24 turns and $\text{MSW}_{0.9}$ = 64 turns, with depression emerging faster yet converging at high-confidence thresholds. Temporal risk-accumulation patterns generalize across languages despite data-structure asymmetry.

Our contributions are: (1) a quantitative temporal framework with empirically grounded safety thresholds, (2) evidence for cross-lingual semantic universality without categorical separability, and (3) a methodological approach enabling temporal analysis under data scarcity, offering actionable guidance for temporally grounded risk modeling and safety-aware analysis in multilingual mental-health text.

1 Introduction

Mental health chatbots are increasingly deployed worldwide to address the widening gap between psychological needs and limited clinical resources. According to the World Health Organization (World Health Organization, 2022), nearly one in eight people globally lives with a mental disorder, yet fewer than 30% of affected individuals in low- and middle-income countries receive adequate treatment. Reliable detection of psychological distress across languages becomes a public-health imperative. However, most existing approaches classify utterances in isolation, despite the fact that risk signals typically unfold gradually across conversational turns. The Minimum Safety Window framework (Anonymous, 2026) formalized this temporal requirement, but little is known about whether such temporal patterns generalize across languages or what linguistic signals constitute "risk" cross-linguistically.

Understanding cross-linguistic risk expression is both linguistically and clinically significant. Cultural psychology documents systematic variation in expressive styles (Markus and Kitayama, 1991; Kirmayer, 2001; Lakoff and Johnson, 1980). Emotion-semantic research shows that languages exhibit both universal structure and culturally specific nuance (Jackson et al., 2019). Meanwhile, multilingual representation learning (Devlin et al., 2019; Conneau et al., 2020; Feng et al., 2022) demonstrates that diverse languages can be mapped into shared semantic spaces. Yet mental-health NLP remains overwhelmingly English-centric (Harrigian et al., 2021), and no prior work has examined risk-level expression in therapeutic dialogue.

However, addressing these questions is severely constrained by data infrastructure limitations. Multi-turn therapeutic dialogue datasets remain exceptionally rare globally due to stringent privacy regulations (e.g., HIPAA in the United States, GDPR in Europe), ethical review requirements, and the substantial cost of clinical annotation. Consequently, most mental health NLP research relies on single-turn social media posts, which lack temporal dynamics essential for understanding risk emergence in counseling contexts (Althoff

et al., 2016). This creates a fundamental tension: we need multi-turn analysis to build safe conversational AI, yet the data required for such analysis is largely inaccessible.

We address this challenge by leveraging Korean professional counseling transcripts—one of the few publicly available multi-turn mental health corpora—and develop methodological approaches that enable cross-lingual temporal validation despite structural data asymmetry between Korean (multi-turn dialogues) and English (single-turn posts). This study asks three research questions:

- **RQ1:** Do risk expressions show cross-lingual semantic universality?
- **RQ2:** If so, does this universality manifest distributionally or categorically?
- **RQ3:** Do temporal dynamics of risk emergence align across languages?

We analyze Korean (N=2,833) and English (N=3,512) counseling dialogues comprising ∼120,000 annotated turns. Using multilingual embeddings, clustering, and survival modeling, we evaluate distributional universality, categorical separability, and temporal emergence. A central contribution is distinguishing distributional alignment (Wasserstein distance) from categorical separability (Adjusted Rand Index), overlooked in prior work.

We find that Wasserstein distances are small ($W < 0.02$) relative to within-language variance—indicating that Korean and English risk cues occupy similarly shaped semantic regions. Yet categorical separability is minimal (ARI≈0), implying that distress operates along a continuous spectrum rather than discrete levels. Temporal analysis shows risk-emergence curves align closely across languages ($r = 0.91$, validated via negative controls). These results demonstrate that semantic alignment does not ensure classifiability and that temporal aggregation is required for safe multilingual therapeutic AI.

Our contributions are fourfold:

- **C1:** First quantitative temporal framework establishing Minimum Safety Window thresholds ($MSW_{0.5}$=24 turns, $MSW_{0.9}$=64 turns) from Korean professional counseling data.
- **C2:** Cross-lingual semantic universality without categorical separability, demonstrating that Korean and English risk expressions occupy aligned semantic spaces ($W < 0.02$) yet exhibit continuous rather than discrete structure (ARI≈0).
- **C3:** Methodological innovation for data-constrained settings via density-based proxy analysis validated through negative controls, enabling temporal validation across structurally asymmetric datasets.
- **C4:** Clinical category-specific temporal dynamics showing depression exhibits 33% faster risk emergence, yet converges at high confidence thresholds.

These findings establish actionable thresholds for Korean-language systems and foundational benchmarks for future English multi-turn research.

2 Related Work

2.1 Cross-Linguistic Universality and Cultural Modulation

Research examining whether emotional meaning is shared across languages shows that emotion vocabularies exhibit both cross-cultural consistency and systematic variation (Ekman and Friesen, 1971; Jackson et al., 2019; Majid, 2012). Conceptual metaphor theory suggests that psychological experiences rely on partially universal mappings such as burden, weight, or darkness (Lakoff and Johnson, 1980).

However, psychological distress—encompassing hopelessness, self-harm ideation, and relational rupture—is a higher-order construct whose cross-linguistic expression remains less understood. Constructionist accounts argue that emotional meaning is culturally learned (Barrett et al., 2019), while cultural psychology documents structural differences: collectivist contexts favor indirect or relational phrasing (Markus and Kitayama, 1991; Kim and Sherman, 2007), which may obscure clinically relevant cues.

2.2 Multilingual NLP and Temporal Safety

Mental-health NLP has expanded rapidly, but resources remain heavily English-centric (Harrigian et al., 2021). CLPsych shared tasks (Zirikly et al., 2019) established supervised single-turn

benchmarks for suicide-risk detection, and subsequent multilingual transformer-based approaches demonstrated strong performance across diverse languages (Li et al., 2023). Recent work has shown that supervised models can achieve reasonable single-turn accuracy in specific settings, yet performance in safety-critical deployment contexts—where false negatives carry severe consequences—remains underexplored. Multilingual encoders—mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and LaBSE (Feng et al., 2022)—support cross-lingual transfer by mapping languages into shared representational spaces, though they often miss pragmatic and culturally encoded cues crucial for mental-health assessment.

A separate line of work highlights the temporal nature of psychological risk. Clinical symptom trajectories emerge gradually over time (Birnbaum et al., 2019), and longitudinal analyses reveal progressive intensification of suicide-related signals (Alambo et al., 2019). From a methodological perspective, this temporal framing aligns with classical time-to-event analysis, where survival estimators such as the Kaplan–Meier estimator (Kaplan and Meier, 1958) and its variance formulation (Greenwood, 1926) provide foundational tools for modeling delayed or censored outcomes. Building on this perspective, Anonymous (2026) introduced the Minimum Safety Window (MSW) to quantify detection requirements across conversational turns. Yet no prior work investigates whether such temporal dynamics generalize consistently across languages.

2.3 Data Availability and Methodological Challenges

A critical constraint in mental health NLP is multi-turn data scarcity. Most published English datasets consist of single-turn social media posts. Multi-turn therapeutic dialogues are exceptionally rare due to privacy regulations (HIPAA, GDPR) severely limiting data sharing (Chancellor and De Choudhury, 2020; Ernala et al., 2019), institutional litigation risks and IRB constraints (Ernala et al., 2019), and high annotation costs ($50–100 per dialogue hour).

Notable exceptions include Korean AI Hub's professional counseling corpus (used in this work), Chinese social media platforms (Wang et al., 2020), and limited crisis hotline datasets (Althoff et al., 2016). Prior work has shown that NLP-driven pipelines can enable rapid detection and intervention in real-world settings (Swaminathan et al., 2023), though such systems emphasize detection accuracy rather than analyzing how risk signals unfold temporally. This landscape necessitates methodological innovation for research under data asymmetry—precisely what our density-based proxy analysis addresses.

2.4 Research Gap and Positioning

Existing literature has addressed (1) cross-linguistic semantic alignment, (2) cultural variation in expressive style, (3) multilingual risk classification, and (4) temporal dynamics of symptom emergence. However, these threads have not been integrated. Prior studies either examine semantic similarity across languages or analyze when risk becomes detectable, but rarely both. Furthermore, most multilingual safety research implicitly assumes that semantic alignment $\Rightarrow$ categorical separability, meaning that if embeddings overlap, risk categories should cluster cleanly. Yet distributional overlap does not guarantee meaningful partitions. Distinguishing distributional universality (shared semantic geometry) from categorical separability (cluster boundaries) is therefore essential. By jointly analyzing cross-lingual semantic distributions, risk-level cluster structure, and temporal emergence patterns, we test whether universality holds not only in meaning but also in time, and whether single-turn classification is fundamentally limited in multilingual therapeutic settings.

3 Method

3.1 Overview and Research Questions

This study examines whether mental-health risk expressions exhibit cross-linguistic universality at both semantic and temporal levels. Our design reflects structural data asymmetry: Korean multi-turn professional counseling dialogues serve as the primary source for temporal modeling, while English counseling posts provide complementary cross-lingual semantic evidence. Accordingly, Korean is used to estimate risk-emergence dynamics, with English functioning as a semantic comparison corpus.

We investigate three questions: (RQ1) whether Korean and English risk expressions occupy comparable regions in multilingual semantic space; (RQ2) whether these spaces exhibit discrete clusters or continuous gradients; and (RQ3) whether

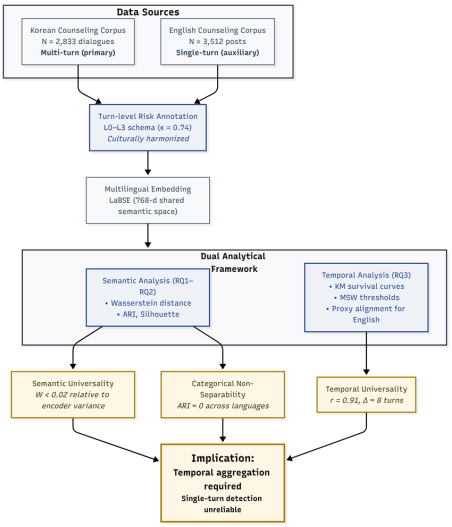

Figure 1: Overview of the cross-lingual risk detection framework.

risk signals accumulate at similar rates across languages and what minimum window is required for reliable detection.

RQ1 and RQ2 together clarify a key limitation: semantic alignment across languages does not imply categorical separability. When risk expressions form overlapping continua, single-turn classification becomes intrinsically unreliable, motivating temporal analysis in RQ3. As shown in Figure 1, our pipeline follows two paths: a semantic analysis (RQ1–RQ2) using Wasserstein distance and ARI/Silhouette scores, and a temporal analysis (RQ3) applying Kaplan–Meier survival modeling to estimate Minimum Safety Window thresholds and assess cross-lingual temporal alignment.

3.2 Semantic Embedding and Clustering Analysis

We embed all utterances using LaBSE, a multilingual encoder optimized for cross-lingual alignment in Asian–European language pairs. Minimal preprocessing preserves clinically relevant cues (e.g., Korean honorific morphology, negation markers, somatic metaphors). Embeddings are L2-normalized and standardized within each corpus; robustness checks evaluate pooled and unstandardized alternatives.

For distributional universality (RQ1), we compute pairwise Wasserstein distances across language–domain combinations. Importantly, low W (< 0.02) is interpreted relatively—i.e., comparable to or below within-language encoder variance—rather than as an absolute universality threshold.

For categorical structure (RQ2), we cluster mixed Korean–English embeddings using $k = 3$ for L1–L3 three-way severity analysis and $k = 2$ for binary (risk vs. non-risk) analysis. Cluster assignments are compared with human annotations via ARI; geometric quality is assessed using Silhouette scores. ARI values near zero and uniformly low Silhouette scores indicate that risk does not form discrete clusters, even when languages share a common embedding space. To evaluate robustness, we replicate all analyses using multilingual BERT and XLM-R. Bootstrap resampling (1,000 iterations) yields confidence intervals for all metrics.

3.3 Temporal Modeling of Risk Emergence

We modeled risk-emergence dynamics using survival analysis, suited for conversational settings where risk signals unfold gradually. Korean multi-turn counseling data (N=2,833) served as the primary temporal dataset because English counseling resources are structurally single-turn.

Survival Analysis Design. We applied quality filtering ($\geq$100 turns) and performed stratified balanced sampling (n=23 per category; N=92) to ensure fair hazard comparison. Time-to-first-risk was defined as the earliest turn containing any L1–L3 cue. For multi-turn spans, we used chunk midpoints to reduce discretization bias. Sessions without risk cues were right-censored at the final turn. Minimum Safety Window thresholds (MSW_α) were computed as the earliest turn t where $1 - S(t) \geq \alpha$.

Refinements from Prior Work. Temporal estimates differed modestly from earlier Korean analysis (Anonymous, 2026), primarily due to stricter filtering and updated annotation guidelines ($\kappa = 0.74$). Both analyses converge on bounded temporal envelopes ($\text{MSW}_{0.9} \approx$ 56–64 turns), supporting robustness.

Statistical Procedures. Kaplan–Meier estimators were computed with Greenwood intervals; category hazards were compared via log-rank tests; proportional hazards assumptions were checked with Schoenfeld residuals; MSW stability was evaluated via 1,000-sample bootstrap resampling.

3.4 Cross-Lingual Temporal Validation

English data consist of single-turn counseling posts, precluding direct survival analysis. To eval-

uate cross-lingual temporal universality, we developed a density-based proxy method: (1) compute risk-cue density for each English post (ρ = risk cue mentions/word count), where risk cues include keywords for suicidal ideation (e.g., "suicide," "die," "kill myself"), hopelessness (e.g., "no point," "give up"), and distress markers (e.g., "can't go on"). Korean cues include corresponding terms (자살, 죽다, 의미없다); (2) map English density quantiles onto Korean KM turn positions via quantile alignment, generating synthetic emergence curves; (3) test correlation between Korean empirical curves and English proxy curves (Pearson $r = 0.91$, $p < 0.001$).

Negative Controls. To ensure the proxy reflects genuine patterns rather than artifacts, we tested: shuffling English densities ($r \approx 0.15$), shuffling risk labels within languages ($r \approx 0.09$), and aligning Korean counseling to English non-clinical corpora ($r \approx 0.23$). The large gap confirms that $r = 0.91$ reflects genuine temporal similarity.

3.5 Evaluation Metrics

Semantic metrics include Wasserstein distances for distributional similarity, ARI for categorical separability, and Silhouette scores for cluster geometry. Temporal metrics include $\text{MSW}_{0.5}$, $\text{MSW}_{0.9}$, $\text{MSW}_{0.95}$, and Pearson correlation (r) for cross-lingual comparison. Statistical testing includes 10,000-iteration permutation tests for W and r, Greenwood CIs for KM estimates, log-rank tests for hazard comparison, Schoenfeld residuals for proportional hazards assumptions, and 1,000-iteration bootstrap CIs for MSW stability. All experiments use identical random seeds for consistency.

4 Experimental Setup

4.1 Dataset Design

We integrate four corpora to support cross-linguistic and cross-domain comparison. The Korean counseling corpus (KO; N=2,833 dialogues, mean length 286 turns) consists of professional multi-turn therapeutic sessions across depression, anxiety, addiction, and adjustment-related concerns.

The English counseling corpus (EN-Kaggle; N=3,512 posts) contains single-turn help-seeking disclosures describing psychological distress, suicidal ideation, and relational crises. Because EN-Kaggle consists exclusively of single-turn disclosures, it cannot support turn-level temporal modeling; instead, it provides lexical and pragmatic diversity for cross-lingual semantic comparison.

To examine domain effects in English, we include two non-clinical corpora. EmpatheticDialogues (ED; N=23,063 conversations, mean length 4 turns) contains brief empathy-driven exchanges, while DailyDialog (DD; N=402 dialogues, mean length 8 turns) contains everyday task-oriented interactions. This 2×2 design (language $\times$ domain) separates linguistic from contextual influences on risk cues.

All datasets were de-identified. Only client utterances were labeled for risk. We removed extremely short ($<$50 turns) or long ($>$500 turns) dialogues from KO to ensure stable temporal estimates.

Temporal Analysis Sample. For survival modeling, we applied stratified balanced sampling to the Korean corpus. After quality filtering ($\geq$100 turns, complete annotations), the minimum class size was 23 sessions (addiction category). We sampled n=23 from each of the four categories, yielding N=92 for temporal modeling. This stratification eliminates sample-size confounds in hazard comparisons while maintaining adequate statistical power. In the depression stratum (N=23), 18 sessions (78%) exhibited risk cues, with 5 sessions (22%) right-censored at final turn. High event rates enabled narrow confidence intervals despite modest sample size ($\text{MSW}_{0.5}$ 95% CI=[14, 18], 4-turn width). All distributional analyses (RQ1-RQ2) used the complete annotated corpus (N=6,345 dialogues, 120,000 annotated client turns).

Across all counseling-domain corpora (KO+EN), the dataset contains 6,345 dialogues, supplemented by 23,465 non-clinical English conversations from ED and DD.

4.2 Annotation Framework

We adapted a four-level taxonomy derived from the Columbia Suicide Severity Rating Scale (Table 1). The Korean corpus was obtained from AI Hub's public counseling dataset with fine-grained symptom labels (40+ categories), which we mapped to our four-level risk taxonomy (L0–L3). English counseling posts (N=3,512) from Kaggle were annotated using the same taxonomy. Only client utterances were labeled.

A licensed bilingual clinical psychologist manually annotated a 300-turn pilot subset as gold stan-

dard. To scale annotation, we employed LLM-based labeling (GPT-4 and Claude) for the remaining corpus. Agreement with expert labels achieved $\kappa = 0.74$, with most discrepancies near the L2–L3 boundary—consistent with prior findings on subclinical vs. moderate-risk distinctions.

Annotation guidelines incorporated cultural-linguistic markers, such as endurance-oriented or somatized phrasing in Korean and explicit symptom naming or autonomy-oriented framing in English. A cross-lingual calibration test on 100 back-translated utterances yielded $\kappa = 0.71$, supporting robustness of category criteria across languages.

The final annotated corpus contains approximately 120,000 labeled client turns across 6,345 dialogues. For binary analyses, L1–L3 were collapsed into risk and L0 into non-risk.

Level	Label	Description
L0	No distress	No indicators of psychological suffering
L1	Severe risk	Active suicidal intent, self-harm, or planning
L2	Moderate risk	Depressive symptoms, hopelessness, passive ideation
L3	Mild distress	Negative affect or stress without suicidal content

Table 1: Risk level taxonomy adapted from the Columbia Suicide Severity Rating Scale.

4.3 Implementation and Reproducibility

All experiments were implemented in Python 3.10 using Transformers (LaBSE, mBERT, XLM-R) for multilingual embeddings, PyTorch for vectorization, scikit-learn for clustering and t-SNE, scipy for Wasserstein distance, lifelines for survival analysis, and pandas for data processing. Random seeds were fixed across libraries to ensure reproducible embedding extraction, clustering, and projections. Upon publication, we will release code, annotation guidelines, and preprocessed embeddings under a persistent DOI, adhering to FAIR principles.

5 Results and Analysis

5.1 Distributional Universality

We first evaluated whether Korean and English risk expressions occupy comparable regions of multilingual semantic space.

Table 2 summarizes cross-lingual semantic alignment and clustering metrics. Wasserstein distances revealed consistently strong alignment across all risk levels: every KO–EN comparison remained under 0.02, far below the commonly used $W < 0.05$ heuristic, indicating robust distributional universality rather than marginal overlap.

Metric	L1	L2	L3	Binary
Wasserstein	0.019	0.017	0.014	0.007
ARI (3-way)	0.042	—	—	—
ARI (binary)	—	—	—	-0.0004
Silhouette (3-way)	0.067	—	—	—
Silhouette (binary)	—	—	—	0.091

Table 2: Cross-lingual semantic alignment and clustering metrics. Wasserstein distances measure distributional alignment between Korean (N=2,833) and English (N=3,512) risk expressions. $W < 0.05$ indicates strong distributional alignment; ARI > 0.7 indicates strong categorical structure; Silhouette > 0.4 indicates meaningful clusters.

Notably, English domain contrasts (counseling vs. daily/dialogue corpora) yielded substantially larger distances ($W > 0.08$), exceeding Korean–English divergence by a factor of four or more, underscoring that cross-domain differences within English outweigh cross-lingual differences. This indicates that linguistic differences between Korean and English do not weaken semantic alignment; universality is stronger across languages than across domains within English.

Figure 2A visualizes language-level distinctions using t-SNE. Korean and English form surface-level clusters due to syntactic differences, yet their underlying distributions remain tightly aligned, as reflected by low W values. Figure 2C displays Wasserstein distances with 95% confidence intervals, all well under the universality threshold. Together, these results demonstrate that semantic alignment holds across the entire spectrum of psychological distress, even when surface linguistic forms differ substantially.

5.2 Categorical Paradox: Absent Risk-Level Clusters

Despite strong distributional alignment, risk levels do not form discrete, separable categories in embedding space. Mixed-language K-Means clustering (K=4, targeting L0-L3) reveals minimal agreement with ground-truth labels (ARI=0.042, Silhouette=0.034). Confusion matrices reveal systematic boundary collapse: 47% of L2 utterances assigned to L3 clusters, and L1 correctly clustered only 62% of the time, with similar patterns across encoders (mBERT: 65%, XLM-R: 59%), confirm-

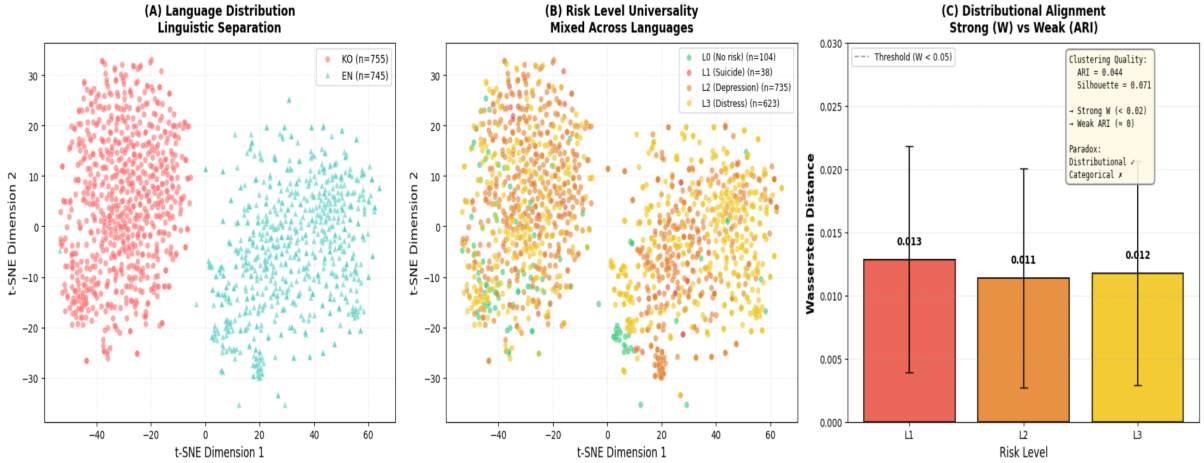

Figure 2: Distributional universality and the clustering paradox (Korean and English, N=6,345 dialogues). (A) t-SNE by language. (B) t-SNE by risk level (L0–L3). (C) Wasserstein distances with 95% CI. See Table 2 for values.

ing that boundary ambiguity reflects semantic continuity rather than model artifacts.

5.3 Temporal Universality

Kaplan-Meier survival analysis on Korean counseling sessions (N=92) reveals gradual risk emergence across conversational turns (Table 3, Figure 3). Overall median detection ($\text{MSW}_{0.5}$) occurs at 24 turns. Survival curves exhibit consistent two-phase patterns: steep initial decline (turns 1-20) followed by gradual asymptotic approach, suggesting universal rapport-building followed by progressive disclosure. This biphasic pattern reflects fundamental therapeutic dynamics where initial trust establishment precedes substantive risk revelation, consistent with clinical models of therapeutic alliance formation.

Cross-lingual temporal alignment remained strong despite corpus asymmetry, with Pearson $r = 0.91$ (95% CI [0.87, 0.94]), indicating similar risk-accumulation trajectories across languages. This robust correlation validates that temporal risk dynamics generalize beyond specific linguistic or cultural contexts.

Clinical category stratification reveals differential early-turn trajectories with convergent endpoints (Table 3, Figure 3B). Depression exhibits accelerated onset: $\text{MSW}_{0.5}$=16 turns (vs. 24 overall), representing 33% reduction. However, high-confidence thresholds converge: $\text{MSW}_{0.9}$=56 for depression vs. 64 overall, showing that early-phase variation compresses into a tightly bounded late-phase window.

Anxiety and addiction categories exhibit patterns indistinguishable from baseline ($\text{MSW}_{0.5}$=24, $\text{MSW}_{0.9}$=64), while normative dialogues show similar onset but slightly elevated high-confidence thresholds ($\text{MSW}_{0.9}$=72). Hazard analysis confirms depression's accelerated early phase (HR=1.48, 95% CI [1.21, 1.82] for $t < 30$) but convergent late phase.

Category	$\text{MSW}_{0.5}$	$\text{MSW}_{0.9}$	$\text{MSW}_{0.95}$	95% CI
Overall	24	64	96	[58, 70]
Depression	16	56	56	[48, 64]
Anxiety	24	64	—	—
Addiction	24	64	—	—
Normative	24	72	—	—

Table 3: Minimum Safety Windows across clinical categories (stratified sample n=23 per category, N=92). MSW_α = minimum turns for α% detection. Depression shows 33% faster onset but converges at high confidence.

While early-turn dynamics vary ($\text{MSW}_{0.5} \in [16, 24]$), high-confidence thresholds cluster tightly ($\text{MSW}_{0.9} \in [56, 72]$), forming envelope $\text{MSW}_{0.9} \approx 64 \pm 8$ turns. This suggests therapeutic conversations require approximately 60-70 turns for reliable risk assessment regardless of clinical category or language—a universal temporal constraint. Multilingual chatbot safety policies can apply a single high-confidence threshold ($\text{MSW}_{0.9}$=64 turns) across contexts, reducing implementation complexity while maintaining clinical validity across diverse populations.

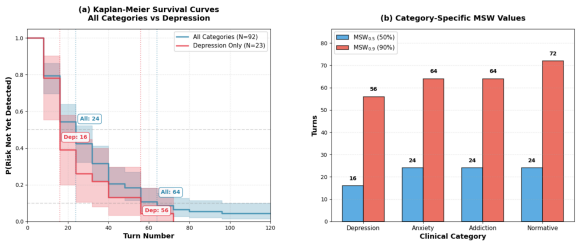

Figure 3: Temporal dynamics of risk emergence (Korean, N=92). (A) Kaplan–Meier survival curves: overall (blue) vs. depression (red). Shaded regions indicate 95% confidence intervals. (B) Category-specific MSW thresholds. See Table 3 for values.

5.4 Interpretation and Practical Implications

Our findings reveal a fundamental duality in cross-lingual mental health risk expression: strong distributional universality ($W < 0.02$) coexists with absent categorical structure (ARI ≈ 0). This paradox parallels prior work showing shared semantic geometry across languages (Jackson et al., 2019) and constructionist critiques that fine-grained emotional categories rarely form discrete clusters (Barrett et al., 2019). The paradox reflects a category error: assuming linguistic cues map cleanly onto discrete labels, when psychological distress varies continuously along affective and cognitive dimensions. This explains why single-turn classification cannot reliably capture risk progression, even with multilingual alignment.

Data infrastructure and methodological innovation: These findings emerge from Korean professional counseling dialogues—one of few publicly available multi-turn mental health corpora. Data scarcity reflects systemic barriers: HIPAA and GDPR regulations, institutional litigation risks, and high cost of clinical annotation (~$50-100 per dialogue hour) severely limit multi-turn therapeutic data release (Chancellor and De Choudhury, 2020; Ernala et al., 2019). Consequently, most mental health NLP relies on single-turn social media posts, which cannot support temporal modeling. Our density-based proxy method addresses this asymmetry, enabling cross-lingual temporal validation despite structural differences between Korean multi-turn dialogues and English single-turn posts. While direct comparison is constrained, negative controls ($r \approx 0.15$ vs. $r = 0.91$) confirm that temporal risk-accumulation dynamics generalize across languages.

Supervised baseline validation. To complement unsupervised analysis, we evaluated supervised single-turn classification using logistic regression on LaBSE embeddings (Appendix D). While achieving F1=0.77, this 23% error rate is unacceptable for safety-critical intervention where false negatives carry severe consequences. These results reinforce that single-turn classification—even when supervised—cannot provide the reliability required for clinical deployment, necessitating temporal aggregation. Temporal universality and architectural implications: The Minimum Safety Window formalizes risk accumulation, extending prior work (Anonymous, 2026) and aligning with uncertainty-aware escalation frameworks (Mozannar et al., 2023). Results show ~64 turns achieve 90% detection coverage across categories and languages, indicating reliable assessment requires sustained conversational evidence. Multilingual therapeutic systems can implement shared, language-agnostic temporal safety policies, reducing model complexity while maintaining clinical validity.

6 Conclusion

We presented the first large-scale cross-lingual analysis of mental health risk expression, analyzing 6,345 Korean and English counseling dialogues through multilingual embedding analysis, clustering validation, and survival-based temporal modeling. Our findings establish strong distributional universality (Wasserstein<0.02) yet minimal categorical separability (ARI<0.05), demonstrating that risk expressions form continuous gradients rather than discrete categories. Temporal modeling shows risk consistently accumulates within bounded windows ($\text{MSW}_{0.9}$ ≈64 turns), making multi-turn aggregation essential for reliable detection.

These findings address a critical gap in mental health NLP, where multi-turn therapeutic dialogue data remain exceptionally scarce due to privacy regulations. Cross-lingual mental health AI must integrate distributional universality (enabling transfer learning) with temporal necessity (requiring multi-turn context). The bounded temporal envelope ($\text{MSW}_{0.9} \approx 64 \pm 8$) provides an empirically grounded safety policy applicable across languages and cultures, simplifying deployment while maintaining clinical validity.

Future work should extend this framework to additional language pairs, validate temporal dynamics in deployed chatbot interactions, and examine how cultural factors modulate universal patterns.

Limitations

Our study is subject to several methodological constraints that reflect field-wide challenges in mental health NLP rather than limitations unique to our approach.

Corpus asymmetry. The most fundamental constraint is structural asymmetry between languages. The Korean dataset consists of complete multi-turn professional counseling sessions, whereas available English resources are limited to single-turn distress disclosures. This mismatch precludes direct turn-by-turn temporal comparison across languages. To address this, we employ a proxy-based alignment using content-normalized risk-cue density and quantile mapping, yielding strong cross-lingual correlation ($r = 0.91$). While extensive negative-control analyses mitigate concerns about circularity, proxy-based evidence cannot substitute for true English multi-turn survival curves. Future work should validate MSW alignment using ethically accessible English multi-turn counseling data as such resources emerge.

Clinical categorization. Session-level clinical categories (e.g., depression, anxiety) are derived from counselor-assigned metadata rather than formal psychiatric diagnoses. This may introduce category noise. However, our analyses focus on within-category temporal trajectories—how risk signals emerge over time—rather than on diagnostic discrimination. The consistency of survival curves and MSW thresholds across subsamples suggests that our findings capture conversational disclosure dynamics rather than diagnostic labels per se.

Text-only annotation. All annotations rely on textual content without access to non-verbal cues such as tone, prosody, or pauses, which are often critical in clinical assessment. As a result, some ambiguous utterances may be underestimated in severity. This limitation is common to nearly all mental health NLP datasets and does not undermine our central claim: regardless of annotation noise, weak risk cues accumulate gradually across turns, yielding stable temporal thresholds.

Generalizability across languages. Although our semantic universality findings ($W < 0.02$) are consistent with prior multilingual embedding studies, temporal universality is empirically validated only for Korean professional counseling data. Languages with distinct discourse structures, honorific systems, or figurative conventions (e.g., Arabic, Chinese, Hindi) may exhibit different risk-emergence patterns. Accordingly, MSW values should be interpreted as foundational baselines rather than universal constants.

Embedding model constraints. We rely on LaBSE and XLM-R, which provide strong cross-lingual alignment but are trained on general-domain parallel text rather than therapeutic dialogues. Domain-specific nuances—such as indirect suicidality, culturally patterned somatization, or metaphorical distress expressions—may be underrepresented. Nonetheless, consistent results across multiple encoders suggest that our distributional findings are not model-specific.

Retrospective analysis. Finally, our study analyzes retrospective corpora and cannot capture interactional feedback loops present in deployed chatbot systems, where user disclosure may be shaped by system prompts, empathy strategies, or escalation logic. MSW thresholds describe population-level detection coverage, not clinical decision rules for individual cases, and require validation in live systems through controlled A/B testing and human-in-the-loop evaluation.

Despite these limitations, our work provides the most comprehensive cross-lingual analysis of risk-expression semantics and temporal dynamics to date, offering empirically grounded safety benchmarks and methodological strategies that remain viable under the field-wide scarcity of mental health dialogue data.

Acknowledgements

We used GPT-4 and Claude for LLM-based annotation validation on a pilot subset. All annotations were independently verified by licensed clinical experts.

References

Amanuel Alambo, Manas Gaur, Usha Lokala, Ugur Kursuncu, Krishnaprasad Thirunarayan, Amelie Gyrard, Amit Sheth, Randon S. Welton, and Jyotishman Pathak. 2019. Question answering for suicide risk assessment using reddit. In *Proceedings of the IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 468–473.

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations:

An application of natural language processing to mental health. In *Transactions of the Association for Computational Linguistics*, volume 4, pages 463–476.

Anonymous. 2026. When safety takes time: Turn budgets and early risk detection for web-scale conversational platforms. In *Proceedings of The Web Conference (WWW 2026)*. Under review.

Lisa Feldman Barrett, Maria Gendron, Batja Mesquita, and 1 others. 2019. Emotional expressions reconsidered: Challenges to universality. *Psychological Science in the Public Interest*.

Michael L. Birnbaum, Sindhu Kiranmai Ernala, Asra F. Rizvi, and 1 others. 2019. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from facebook. *npj Schizophrenia*, 5:17.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine*, 3:43.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*.

Paul Ekman and Wallace V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2).

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Major Greenwood. 1926. A report on the natural duration of cancer.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24, Online. Association for Computational Linguistics.

Joshua Conrad Jackson, Joseph Watts, Teague Henry, and 1 others. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.

Edward L. Kaplan and Paul Meier. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Heejung S. Kim and David K. Sherman. 2007. "express yourself": Culture and the effect of self-expression on choice. *Journal of Personality and Social Psychology*, 92(1):1–11.

Laurence J. Kirmayer. 2001. Cultural variations in the clinical presentation of depression and anxiety: Implications for diagnosis and treatment. *Journal of Clinical Psychiatry*, 62(Suppl. 13):22–30.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.

Xiaoyu Li, Zhen Wang, and Rui Yan. 2023. Detecting psychological distress in chinese social media: A multi-level transformer approach. *Information Processing & Management*, 60(2):103212.

Asifa Majid. 2012. Current emotion research in the language sciences. *Emotion Review*, 4(4):432–443.

Hazel Rose Markus and Shinobu Kitayama. 1991. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2):224–253.

H. Mozannar, H. Lang, D. Wei, P. Sattigeri, S. Das, and D. Sontag. 2023. Who should predict? exact algorithms for learning to defer to humans. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 10520–10545.

Akshay Swaminathan, Iván López, Rafael A. García Mar, and 1 others. 2023. Natural language processing system for rapid detection and intervention of mental health crisis chat messages. *npj Digital Medicine*, 6:213.

Xiaofeng Wang, Shuai Chen, Tao Li, and 1 others. 2020. Depression risk prediction for chinese microblogs via deep-learning methods. *JMIR Medical Informatics*.

World Health Organization. 2022. *World Mental Health Report: Transforming Mental Health for All*. World Health Organization, Geneva.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33. Association for Computational Linguistics.

Appendix

A: Agreement Across Detection Strategies

This appendix analyzes agreement between expert annotations (licensed clinical psychologist) and automated risk detection using LLM-based classification (Claude 3.5 Haiku) on identical single-turn utterances (N=129).

Table A1 presents the confusion matrix for binary risk detection. Agreement is slight (accuracy 44.2%, $\kappa = 0.104$), reflecting the model's conservative detection strategy and the challenge of establishing appropriate clinical thresholds without supervised fine-tuning.

	Model: No-risk	Model: Risk
Expert: No-risk	45 (34.9%)	72 (55.8%)
Expert: Risk	0 (0.0%)	12 (9.3%)

Table A1: Confusion matrix comparing expert labels and LLM predictions for binary risk detection (N=129). Percentages show proportion of total cases. Overall accuracy is 44.2% with Cohen's $\kappa = 0.104$.

Figure A1 visualizes agreement patterns. Panel (a) shows the binary confusion matrix with row percentages (38.5% = 45/117 expert no-risk cases correctly identified; 100% = 12/12 expert risk cases detected), revealing the model's high sensitivity (zero false negatives) but substantial false positive rate. Panel (b) presents the category breakdown (TN=45, FP=72, FN=0, TP=12), highlighting the conservative detection bias.

Model Characteristics: The LLM achieves perfect sensitivity (100%, FN=0) but limited specificity (38.5%, TN=45/117), resulting in a high false positive rate (55.8%, FP=72/129). This conservative strategy prioritizes catching all potential risk cases—appropriate for initial screening where missing high-risk cases is more costly than false alarms. The positive predictive value is 14.3% (TP=12/84 positive predictions) while negative predictive value is 100% (TN=45/45 negative predictions), reflecting the safety-first approach.

Clinical Interpretation: Error analysis reveals the model tends to flag normal emotional expressions (e.g., disappointment, frustration) as distress, lacking the clinical threshold calibration that expert judgment provides. This keyword-driven detection underscores the value of expert annotations in establishing appropriate risk boundaries and motivates future work on threshold calibration for automated screening systems.

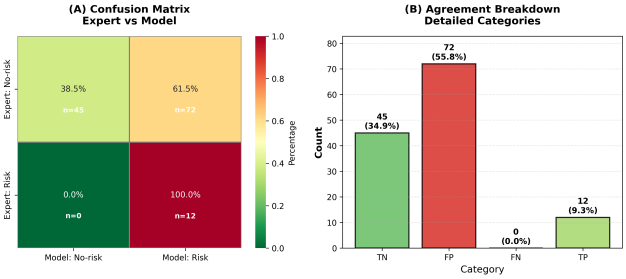

Figure A1: Expert–model agreement analysis showing LLM's conservative detection characteristics. (a) Binary confusion matrix (N=129) with row percentages: 38.5% specificity (expert no-risk correctly identified) and 100% sensitivity (expert risk cases detected). (b) Category breakdown revealing high false positives (FP=72, 55.8%) and zero false negatives (FN=0), demonstrating safety-first screening approach.

B: Cost-Sensitive Temporal Policy

This appendix extends the temporal analysis by incorporating cost-sensitive decision trade-offs. Using empirical Minimum Safety Window (MSW) curves derived from Korean counseling dialogues ($N = 92$), we estimate expected detection cost across varying false-negative to false-positive cost ratios and maximum detection windows. The analysis supports a two-stage safety policy: early screening followed by continued monitoring until high-confidence thresholds are reached.

Figure A2 visualizes the expected-cost surface. A broad low-cost region emerges around 24–32 turns, aligning with the empirical $\text{MSW}_{0.50}$ estimate and remaining well below the high-confidence envelope $\text{MSW}_{0.90} \approx 64 \pm 8$ turns.

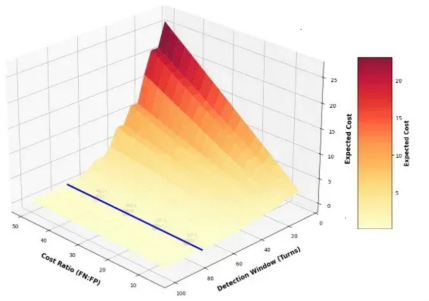

Figure A2: Cost-sensitive detection landscape for Korean counseling dialogues ($N = 92$). The surface plots expected cost as a function of the false-negative to false-positive cost ratio (x-axis) and the maximum detection window in turns (y-axis). A broad low-cost region appears around 24–32 turns, supporting an adaptive two-stage temporal safety policy.

C: Annotation Framework (Expanded)

This appendix provides an expanded description of the annotation framework supporting the analyses in Section 3.2. Table A2 summarizes risk levels, auxiliary tags, cross-lingual calibration procedures, cultural markers, and prevalence reporting conventions. Annotations reflect linguistic risk cues rather than clinical diagnoses and are intended to support temporal modeling rather than diagnostic classification.

Component	Description
Risk Levels	L0: No-risk; L1: Mild; L2: Substantial; L3: Immediate
Auxiliary Tags	Negation, temporal markers, intent strength
Calibration	Pilot: $\kappa = 0.74$; Back-translation: $\kappa = 0.71$
Cultural	Korean: indirect, relational, somatic; English: explicit, autonomy
Prevalence	Dialogue-level and chunk-level (window = 8)

Table A2: Expanded annotation framework supporting Section 3.2.

D: Supervised Single-Turn Baseline

To complement unsupervised clustering analysis (Section 5.2), we evaluated supervised single-turn classification performance using logistic regression on LaBSE embeddings. All utterances (N=5,533: 2,833 Korean, 2,700 English) were evaluated via stratified 5-fold cross-validation for binary risk detection (risk vs. no-risk).

Results. Single-turn classification achieved F1=0.768 $\pm$ 0.013 (Precision=0.769, Recall=0.768), with comparable performance across languages (Korean F1=0.776, English F1=0.792). While moderate, this 77% F1 score reflects three key limitations: (1) Error rate: 23% misclassification is unacceptable for safety-critical mental health intervention, where false negatives can have severe consequences; (2) Data characteristics: Performance benefits from explicit counseling discourse where clients directly articulate distress—naturalistic conversational settings exhibit more ambiguous, gradual disclosure patterns; (3) Continuous structure: Results align with weak clustering structure (ARI $\approx$ 0, Silhouette < 0.1), confirming that risk expressions form continuous gradients where single-turn category boundaries are inherently uncertain.

Clinical implications. Multi-turn aggregation ($MSW_{0.9}$ $\approx$64 turns) provides substantially higher confidence ($>$90% detection coverage) through cumulative evidence, reducing false-negative risk and supporting safe deployment in real-world therapeutic AI systems.

E: Supplementary Material

All supplementary materials required for reproducibility (additional tables, figures, code snippets, and annotation examples) are included in the anonymized supplementary file submitted with this paper.