Supplementary File: ProtoSiTex: Learning Semi-Interpretable Prototypes for Multi-label Text Classification

U. K. Nareti, S. Kumar, S. Pandey, S. Chattopadhyay, C. Adak,

Appendix A
Computational Complexity Analysis

In Table A.1, we list the symbols used in the complexity analysis of ProtoSiTex. The time and space complexities of various modules and phases are mentioned below.

1) Text Encoder: Let $C_{\text{enc}}(t,d)$ denote the cost of the language-model encoder. If the encoder is frozen, this cost is incurred only once per epoch. The encoder stores $\mathcal{E} \in \mathbb{R}^{t\times d}$, requiring $\mathcal{O}(td)$ memory.

2) Prototype Projection via Multi-Head Attention (MHA): The complexities for MHA are:

$$\text{Time: } \mathcal{O}\big(h(qd^2 + 2td^2 + 2qtd) + qd^2\big);$$

$$\text{Space: } \mathcal{O}(td + qd + hqt).$$

If the standard shared-projection implementation of MHA is used (head dimension $d_h = d/h$), the cost simplifies to $\mathcal{O}\big((t+q)d^2 + 2qtd\big)$.

3) Clustering Phase: During prototype discovery, cosine similarities between subsentence embeddings and prototypes are computed to form $\mathcal{X}{=}\mathcal{P}{\cdot}\mathcal{E}^\top$. The proximity and regularization losses require pairwise distances and orthogonality among prototypes:

$$\text{Time: } \mathcal{O}(tqd + q^2 d), \quad \text{without MHA},$$

$$\text{Time: } \mathcal{O}\big(tqd + q^2 d + \text{MHA}(t,q,d,h)\big), \quad \text{with MHA},$$

$$\text{Space: } \mathcal{O}(td + qd + hqt).$$

4) Classification Phase: The classification phase maps prototype representations to labels and performs hierarchical aggregation. Let $\tilde{\mathcal{G}}_1 = f(\mathcal{H})$ denote the MLP-based proto-to-label predictor and $\tilde{\mathcal{G}}_2$, $\tilde{\mathcal{G}}_3$ the sentence- and document-level outputs.

$$\text{Time: } \mathcal{O}(qdl + tqd + tql + ntl) + \text{MHA}(t,q,d,h),$$
$$\text{Space: } \mathcal{O}(td + qd + hqt).$$

The first term arises from dense mappings, while the remaining terms originate from prototype–subsentence assignment and hierarchical aggregation via $\mathcal{M}_2$ and $\mathcal{M}_1$.

5) Per-Epoch Complexity: Combining both phases, the per-epoch training cost is:

$$\mathcal{O}\Big(i_1\big[C_{\text{enc}} + tqd + q^2 d + \text{MHA}(t,q,d,h)\big] +$$
$$i_2\big[C_{\text{enc}} + qdl + tqd + tql + ntl + \text{MHA}(t,q,d,h)\big]\Big),$$

with peak memory $\mathcal{O}(td + qd + hqt)$. In practice, the terms proportional to $(t+q)d^2$ from MHA and C_{enc} dominate computation, while attention maps (hqt) dominate memory. The hierarchical aggregation adds $\mathcal{O}(ntl)$, which remains minor for moderate l. Hence, the overall complexities of the two phases can be summarized as:

$$\text{Clustering: } \mathcal{O}(tqd + q^2 d),$$

$$\text{Classification: } \mathcal{O}(qdl + tqd + tql + ntl),$$

with an additional $\mathcal{O}\big((t+q)d^2 + 2qtd\big)$ factor when multi-head attention is applied.

6) Discussion on Scalability and Stability: The above analysis highlights that the computational cost of ProtoSiTex grows linearly with both the number of subsentences t and prototypes q, while quadratically with the embedding dimension d. Since q and h are typically small ($<$64 and $<$8), the model remains scalable to long documents and large vocabularies. Moreover, the alternating dual-phase training decomposes the optimization into tractable sub-problems, ensuring bounded per-iteration cost and stable convergence. Empirically, we observed smooth loss reduction across epochs, demonstrating that the hierarchical coupling of clustering and classification does not introduce training instability.

TABLE A.1: Symbols used in complexity analysis

Symbol	Description
t	Number of subsentences in a document
n	Number of sentences in a document
q	Number of learnable prototypes
d	Dimensionality of embedding space
h	Number of attention heads
l	Number of output labels or classes
i_1, i_2	Iterations of clustering and classification phases per epoch
$C_{\text{enc}}(t,d)$	Computational cost of the language-model encoder
$\mathcal{M}_1, \mathcal{M}_2$	Document-to-sentence and sentence-to-subsentence mappings
$\mathcal{P}$	Prototype embeddings ($\mathcal{P} \in \mathbb{R}^{q\times d}$)
$\mathcal{E}$	Subsentence embeddings ($\mathcal{E} \in \mathbb{R}^{t\times d}$)
$\mathcal{H}$	Prototype-aware representations after MHA ($\mathcal{H} \in \mathbb{R}^{q\times d}$)
$\tilde{\mathcal{G}}_1, \tilde{\mathcal{G}}_2, \tilde{\mathcal{G}}_3$	Predicted labels at prototype, sentence, and document levels

Appendix B
Prototype Update Stability and Effect of Loss Weights

1) Prototype Update Stability: The prototypes $\mathcal{P} = \{p_1, p_2, \ldots, p_q\}$; $\forall p_i \in \mathbb{R}^d$ are updated exclusively during the clustering phase, while the classifier parameters remain frozen. This alternate design decouples prototype discovery from label prediction and prevents gradient interference between representation and supervision. Let $\theta_{\mathcal{P}}$ and θ_C denote the parameters of the prototype and classification modules, respectively. The alternate optimization follows:

$$\theta_{\mathcal{P}}^{(k+1)} \leftarrow \theta_{\mathcal{P}}^{(k)} - \eta_{\mathcal{P}} \nabla_{\theta_{\mathcal{P}}} \mathcal{L}_{cr},$$
$$\theta_C^{(k+1)} \leftarrow \theta_C^{(k)} - \eta_C \nabla_{\theta_C} \mathcal{L}_{cs}, \tag{1}$$

where, $\eta_{\mathcal{P}}$ and η_C are phase-specific learning rates. Since $\mathcal{L}_{cr}$ is dominated by smooth cosine and Euclidean similarity terms, the gradient field $\nabla_{\theta_{\mathcal{P}}} \mathcal{L}_{cr}$ remains Lipschitz-continuous, ensuring bounded step sizes under standard $\eta_{\mathcal{P}} < 2/\mathcal{L}_{cr}$ conditions. The diversity regularization further introduces a repulsive term between prototypes, reducing the risk of prototype collapse and stabilizing updates. Empirically, the alternate schedule ($i_1{=}10$, $i_2{=}100$) exhibited monotonic

decline of both $\mathcal{L}_{cr}$ and $\mathcal{L}_{cs}$, indicating that prototype refinement converges before each classification update and avoids oscillatory behavior. Consequently, the dual-phase training can be viewed as a bounded coordinate-descent process over $(\theta_{\mathcal{P}}, \theta_C)$ that converges to a stationary solution of the joint objective $\mathcal{L}_{joint} = \mathcal{L}_{cr} + \mathcal{L}_{cs}$.

2) Effect of α and λ Weighting: The coefficients $\alpha = \{\alpha_1, \alpha_2, \alpha_3\}$ and $\lambda = \{\lambda_1, \lambda_2, \lambda_3\}$ balance the contributions of different loss components in the clustering and classification phases:

$$\begin{aligned} \mathcal{L}_{cr} &= \alpha_1 \mathcal{L}_{prox} + \alpha_2 \mathcal{L}_{reg} + \alpha_3 \mathcal{L}_{cs}, \\ \mathcal{L}_{cs} &= \lambda_1 \mathcal{L}_1(\mathcal{G}_1, \tilde{\mathcal{G}}_1) + \lambda_2 \mathcal{L}_2(\mathcal{G}_2, \tilde{\mathcal{G}}_2) + \lambda_3 \mathcal{L}_3(\mathcal{G}_3, \tilde{\mathcal{G}}_3), \\ \sum_i \alpha_i &= \sum_i \lambda_i = 1. \end{aligned} \tag{2}$$

Here, α_1 controls prototype–data proximity, encouraging semantic alignment. Increasing α_1 accelerates convergence but may reduce diversity. α_2 governs regularization, promoting sparsity and orthogonality; higher α_2 values improve prototype separation, but may slow adaptation. The coupling term α_3 weights the auxiliary classification feedback and stabilizes prototypes against noisy subsentence embeddings. A small non-zero α_3 (e.g., 0.01) ensures mild supervision without overriding unsupervised clustering, yielding a smooth prototype trajectories. Here, λ_1, λ_2, and λ_3 control hierarchical supervision across prototype-, sentence-, and document-level predictions. Large λ_1 values tighten prototype–label alignment, enhancing fine-grained interpretability but slightly increasing training variance. Conversely, a higher λ_3 emphasizes global consistency and accelerates convergence on document-level labels. Empirically, on HR, the configuration $(\alpha_1, \alpha_2, \alpha_3)$=$(0.01, 0.98, 0.01)$ and $(\lambda_1, \lambda_2, \lambda_3)$=$(0.6, 0.2, 0.2)$ provided the best trade-off between interpretability, stability, and accuracy (Fig. 4 of the main paper). Within this range, the loss landscape remained well-conditioned, and gradient norms across epochs exhibited bounded variance, confirming the robustness of the weighting scheme.

Overall, the alternating optimization with balanced α and λ weighting yields stable prototype evolution and smooth loss trajectories. The independence of clustering and classification gradients ensures that the model avoids catastrophic drift of prototypes, resulting in consistent semantic alignment and reproducible convergence across runs.

Appendix C
Convergence and Optimization Behavior

Fig.s C.1 and C.2 jointly provide comprehensive evidence of the stability and optimization behavior of ProtoSiTex. Fig. C.1 reports the evolution of all loss components after every epoch of the clustering and classification phases, whereas Fig. C.2 averages the same metrics after one complete alternation cycle (clustering $\curvearrowright$ classification). Together, these plots highlight both the fine-grained intra-phase and averaged inter-phase convergence dynamics of the proposed model. Across both figures, all loss components exhibit a smooth, monotonic reduction, confirming that the alternate optimization effectively minimizes a tight upper bound of the joint objective, and converges toward a stationary solution.

$$\mathcal{L}_{joint} = \mathcal{L}_{cr} + \mathcal{L}_{cs}, \tag{3}$$

1) Clustering-Phase Stability: The interpretability loss $\mathcal{L}_{int}$, prototype-matching loss $\mathcal{L}_{pm}$, and proximity loss $\mathcal{L}_{prox}$ (Fig.s C.1–C.2) display an early sharp decrease followed by gradual flattening, indicating that prototypes become semantically coherent and well-aligned in the latent space. Simultaneously, the sparsity loss $\mathcal{L}_s$ and diversity loss $\mathcal{L}_d$ decrease steadily, enforcing compact yet mutually dissimilar prototypes. Their combination, expressed as the structural regularization loss $\mathcal{L}_{reg}$ declines monotonically, ensuring orthogonality and preventing prototype collapse.

2) Classification-Phase Stability: The prototype-to-label ($\mathcal{L}_1$), sentence-level ($\mathcal{L}_2$), and document-level ($\mathcal{L}_3$) losses show rapid initial drops followed by smooth stabilization, implying efficient hierarchical knowledge transfer from prototypes to higher-level representations. Both the classification loss $\mathcal{L}_{cs}$ and clustering loss $\mathcal{L}_{cr}$ converge quickly, demonstrating that alternate training maintains a balanced optimization between discriminative and structural objectives.

3) Global Convergence and Equilibrium: The overall joint loss $\mathcal{L}_{joint}$ decreases exponentially before stabilizing, indicating that further iterations yield marginal improvements. This behavior empirically supports the theoretical view that the alternating procedure acts as a block-coordinate descent minimizing an upper bound of $\mathcal{L}_{joint}$.

Under standard smoothness and boundedness conditions, Lipschitz-continuous gradients of $\mathcal{L}_{prox}$ and $\mathcal{L}_{reg}$, compact parameter sets, and bounded step sizes $\eta_{\mathcal{P}} < 2/\mathcal{L}_{cr}$ and $\eta_C < 2/\mathcal{L}_{cs}$, each phase ensures non-increasing sub-objectives. Empirical results in Fig.s C.1–C.2 exhibit no oscillations or divergence, validating these assumptions. Moreover, adopting an alternate ratio of $10 : 100$ $(= i_1 : i_2)$ guarantees that prototypes reach local equilibrium before classifier updates, thus maintaining approximate gradient orthogonality between $\nabla_{\theta_{\mathcal{P}}} \mathcal{L}_{cr}$ and $\nabla_{\theta_C} \mathcal{L}_{cs}$.

The synchronized decay of all loss components confirms that ProtoSiTex attains stable and reproducible optimization behavior. The monotonic convergence of interpretability, sparsity, proximity, and predictive objectives demonstrates that the model achieves a well-conditioned multi-objective balance.

Appendix D
Prototype Interpretability Measurement

This section presents the interpretability evaluation. ProtoSiTex is evaluated using a set of metrics that quantify how effectively the learned prototypes represent the data, distinguish between classes, align with the latent space, and faithfully reflect ProtoSiTex's decision-making process.

1) Coverage (ξ): It measures the representativeness of the prototype set, ensuring that every region of the latent space has at least one nearby prototype [1]. It quantifies how comprehensively the prototypes capture the overall data distribution and semantic variability. A high ξ thus

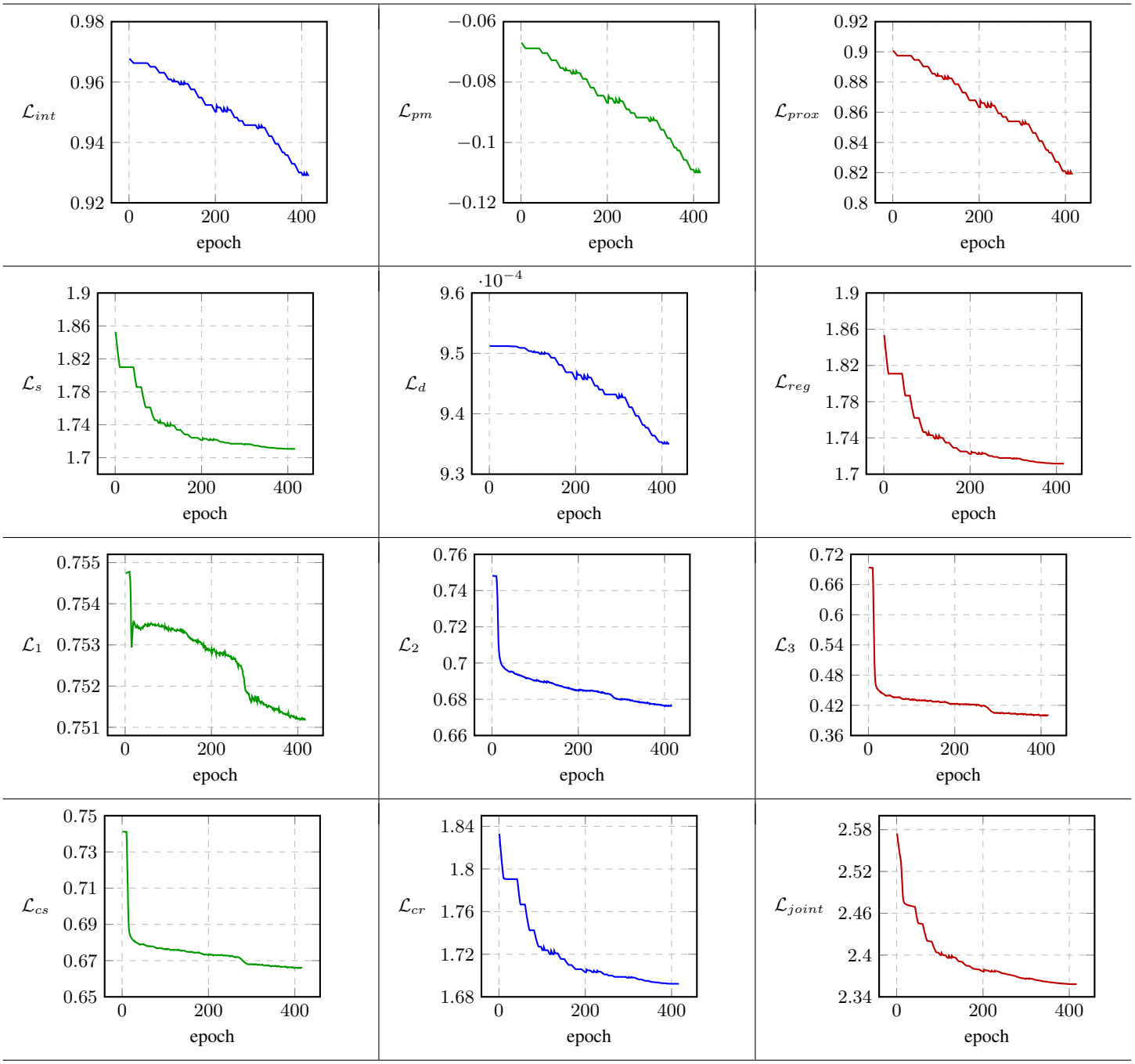


Fig. C.1: Convergence analysis (loss is calculated after every epoch of clustering and classification phase)

indicates that the learned prototypes provide a complete and non-redundant representation of the latent space, supporting comprehensive and interpretable decision explanations. The document-level coverage is defined as:

$$\xi_{doc} = \frac{1}{t}\sum_{i=1}^{t} \frac{\sum\limits_{j\in\mathscr{L}_i} \max\limits_{p_k\in\mathcal{P}_j} \phi\left(\frac{\mathcal{E}_i^\top p_k}{||\mathcal{E}_i||_2||p_k||_2}\right)}{\max\{1,|\mathscr{L}_i|\}} \tag{4}$$

where, $\phi(x) = (x+1)/2$, t is the number of subsentences in the document, $\mathscr{L}_i$ is the set of active labels at subsentence i, and $\mathcal{P}_j$ the prototype set assigned to label j. Dataset-level coverage is obtained by averaging ξ_{doc} over all documents.

$$\xi = \frac{1}{M}\sum_{d=1}^{M} \xi_{doc}^{<d>}, \tag{5}$$

The inner $\max_{p_k\in\mathcal{P}_j}$ in Eqn. (4) captures the strongest prototype–embedding alignment in the cosine space for each active label j, while normalization by $\max\{1,|\mathscr{L}_i|\}$ accounts for multi-label balance and prevents division by zero. The mapping $\phi(x) = (x+1)/2$ bounds all similarity scores within $[0,1]$, ensuring that ξ_{doc} and ξ remain normalized and comparable across documents. The metric attains $\xi =$

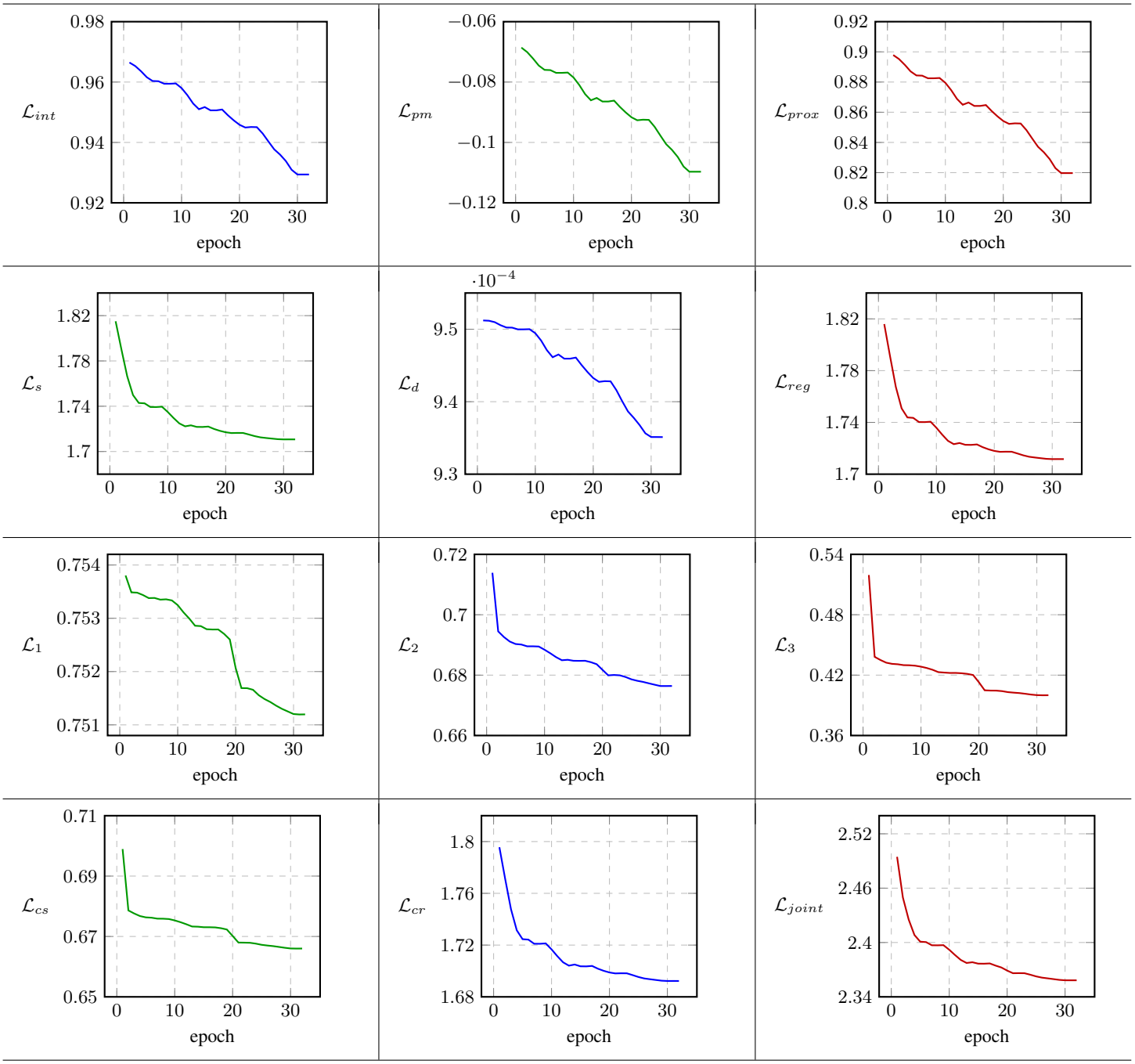


Fig. C.2: Convergence analysis (mean loss is calculated after every complete alternate cycle)

1 only when every active label at every subsentence is perfectly represented by at least one aligned prototype, thereby providing a consistent measure of representational completeness.

2) Contrastivity (ζ): This quantifies the discriminative power of the prototype set, measuring how well the prototypes are separated in the latent space [2]. While coverage captures representational completeness, contrastivity reflects representational exclusivity, ensuring that prototypes of different classes remain distinct and non-overlapping. Thus, a high ζ indicates greater inter-prototype diversity and stronger class separation.

The metric ζ is computed as the mean pairwise cosine dissimilarity between normalized prototype embeddings:

$$\zeta = \frac{1}{q(q-1)} \sum_{\substack{\mathcal{P}_i, \mathcal{P}_j \in \mathcal{P}; \\ i \neq j}} \left[1 - \phi \left(\frac{\mathcal{P}_i^\top \mathcal{P}_j}{||\mathcal{P}_i||_2, ||\mathcal{P}_j||_2} \right) \right] \quad (6)$$

where, $\phi(x) = (x+1)/2$ maps the cosine similarity from $[-1, 1]$ to $[0, 1]$ for bounded dissimilarity. Each term in Eqn. (6) measures the pairwise orthogonality between prototypes, penalizing redundant or collinear representations. Normalization by $q(q-1)$ ensures scale invariance between different prototype counts, and bounded $\phi(\cdot)$ transformation

guaranties $\zeta \in [0, 1]$. A higher value of ζ indicates a more uniformly dispersed prototype set, implying a more eloquent discriminative structure and reduced redundancy in the learned latent space.

3) Centered Kernel Alignment (κ): It quantifies the structural correspondence between data embeddings and learned prototypes in the latent space [3]. While ξ and ζ measure representational completeness and exclusivity, respectively, κ evaluates relational consistency, i.e., how well the pairwise relationships among data embeddings are preserved by the prototypes. A high κ indicates that prototype geometry faithfully mirrors the latent structure of the input space. The linear kernel–based κ is defined as:

$$\kappa = \frac{||\tilde{\mathcal{E}}\tilde{\mathcal{P}}^\top||_F^2}{||\tilde{\mathcal{E}}\tilde{\mathcal{E}}^\top||_F||\tilde{\mathcal{P}}\tilde{\mathcal{P}}^\top||_F} \tag{7}$$

where, $\tilde{\mathcal{E}}$ and $\tilde{\mathcal{P}}$ are mean-centered embeddings of all subsentences and prototypes, respectively, and $||\cdot||_F$ denotes the Frobenius norm. Eqn. (7) computes the normalized alignment between data– and prototype–kernel spaces, yielding $\kappa \in [0, 1]$. Mean-centering eliminates first-order bias, ensuring that κ reflects only second-order relational structure. A higher κ value implies the prototype similarity matrix is well aligned with that of the data, confirming prototypes preserve the semantic geometry of the latent representation space.

4) Fidelity (φ): This measures the faithfulness of the explanations based on prototypes [2], i.e., how the prototypes accurately reproduce the behavior of the model's own decision. While ξ and ζ assess representational structure, φ evaluates behavioral consistency between model predictions and their prototype-derived counterparts. A high value of φ indicates that the prototypes genuinely drive the reasoning process of the network. Let $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$ be the set of m documents with corresponding prototype subsets $\mathcal{P}_d = \{\mathcal{P}_{d_1}, \ldots, \mathcal{P}_{d_m}\}$. For each document, the model produces latent-space predictions on both the original data and its prototypes, denoted as $\mathcal{Y}_d = \{y_1^d, \ldots, y_m^d\}$ and $\mathcal{Y}_p = \{y_1^p, \ldots, y_m^p\}$, respectively. Using $\mathcal{Y}_d$ as a pseudo ground truth, the fidelity scores are then computed using three evaluation measures: macro-averaged F1 ($\varphi_{\mathcal{F}_m}$), balanced accuracy ($\varphi_{\mathcal{BA}_m}$), and complementary Hamming loss ($\varphi_{\mathcal{HL}^c}$). Fidelity φ captures the functional alignment between the model and its interpretable prototype subspace. The averaging across complementary metrics ensures stability and penalizes overfitting to any single criterion. Since $\varphi \in [0, 1]$, higher values indicate greater predictive agreement, demonstrating that the learned prototypes provide faithful, semantically coherent, and behaviorally consistent explanations of the model's decision process.

The prototype interpretability metrics (Fig. D.1) indicate a well-balanced solution across representational completeness, separation, relational consistency, and behavioral faithfulness. In ProtoSiTex, *first*, the coverage score ξ equals to 0.69 suggests that the learned prototypes collectively span a substantial portion of the latent manifold.The model, therefore, maintains a broad representational reach across multi-label factors. *Second*, contrastivity ζ=0.394 reflects moderate inter-

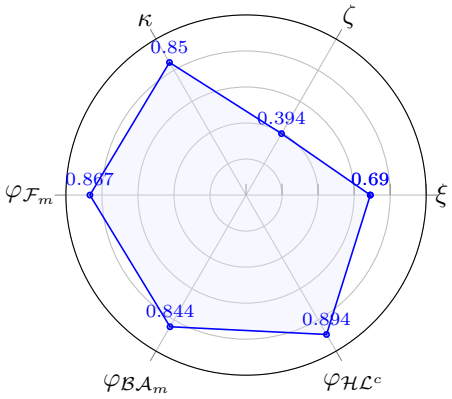


Fig. D.1: Prototype interpretability metrics of ProtoSiTex

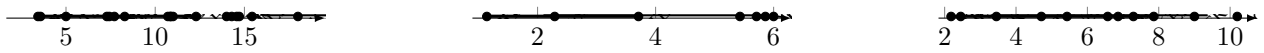
prototype separation under cosine dissimilarity. In conjunction with ξ=0.69, this level of ζ is consistent with a non-redundant prototype set, which avoids collapse while not over-dispersing prototypes to the point of sacrificing coverage (the typical coverage–separation trade-off). *Third*, the centered kernel alignment κ=0.850 evidences strong relational consistency between the prototype and data similarity structures after mean-centering, i.e., second-order geometry of the data is well preserved by the prototypes. This is important because high κ complements ξ and ζ, whereas ξ and ζ quantify first-order coverage and pairwise separation; κ certifies that the pattern of relationships among samples is mirrored by the prototype space. Finally, fidelity (behavioral faithfulness) is high across all three complementary measures, with $\varphi_{\mathcal{F}_m}$=0.867, $\varphi_{\mathcal{BA}_m}$=0.844, and $\varphi_{\mathcal{HL}^c}$=0.894. Since each metric is bounded in $[0, 1]$, these values indicate strong agreement between the model's predictions on real data and prototype-derived predictions.

Taken together, the tuple $(\xi, \zeta, \kappa, \varphi_{\mathcal{F}_m}, \varphi_{\mathcal{BA}_m}, \varphi_{\mathcal{HL}^c})$ characterizes a regime with (i) substantial manifold coverage, (ii) sufficient separation to prevent redundancy, (iii) high alignment of prototype and data geometry, and (iv) strong behavioral consistency. This aligns with the theoretical objective of ProtoSiTex, jointly optimizing prototype quality and hierarchical prediction, by demonstrating that the learned prototypes are semantically representative, structurally well-organized, and functionally faithful.

APPENDIX E
STATISTICAL SIGNIFICANCE TEST

To assess the reliability of the observed performance differences, we performed a non-parametric statistical significance analysis following [4] and [5]. The Friedman test with the Iman–Davenport correction was used to verify global differences among competing models, followed by the Nemenyi post-hoc procedure to identify pairwise significance at α=0.05.

1) All Model Comparison: A comprehensive nonparametric statistical analysis was conducted across all k=18 models, encompassing both black-box and prototype-based architectures (Table III of the main paper), and evaluated



(a) All models (k=18), CD = 9.46 (b) Prototype models (k=7), CD = 3.83 (c) Black-box models (k=11), CD = 5.87

Fig. D.2: Critical Difference (CD) diagrams for Friedman–Nemenyi analysis. Points show mean ranks; horizontal bars connect models that are not significantly different at α=0.05.

on N=7 dataset–metric pairs (i.e., HR: $\mathcal{F}_m, \mathcal{BA}_m, \mathcal{HL}^c$; IMDb: $\mathcal{A}, \mathcal{F}$; TweetEVAL: $\mathcal{A}, \mathcal{F}$). The Friedman test yielded a chi-square statistic of χ_F^2=95.98, quantifying the deviation of model performance ranks from the null hypothesis of equivalent behavior. To improve the accuracy for finite samples, the Iman–Davenport adjustment was applied, resulting in an F-statistic of F_F=25.0 ($p \ll 10^{-6}$). These values indicate that the differences among the competing models are statistically significant, thereby rejecting the null hypothesis of equal performance with high confidence. The Nemenyi critical difference (CD) of 9.46 was used for pairwise comparisons of the average ranks. ProtoSiTex achieved the lowest mean rank ($\bar{R}$=2.00), followed closely by RoBERTa-Large (2.29) and DeBERTaV3 (4.00). The differences between ProtoSiTex and the strongest black-box transformers were within the CD threshold, signifying no significant degradation relative to these SOTA models. However, ProtoSiTex significantly outperformed major prototype-based architectures (e.g., Proto-LM, ProtoryNet, ProSeNet, ProtoLens, ProtoTEX) and legacy models (e.g., ALBERT, TextGCN), whose rank differences exceeded the CD. This confirms that ProtoSiTex achieves statistically superior generalization among interpretable frameworks while remaining competitive with the best-performing black-box baselines.

2) Prototype-based Comparison: Restricting the analysis to prototype-driven interpretable models (k=7), the Friedman statistic yielded χ_F^2=17.0 and F_F=4.08 (p=0.0028), rejecting the null hypothesis of equal performance. With a CD=3.83, ProtoSiTex ($\bar{R}$=1.14) was found to be significantly better than Proto-LM, ProtoryNet, ProSeNet, and ProtoLens, while their differences with GAProtoNet ($\bar{R}$=2.29) and ProtoTEX ($\bar{R}$=3.71) remained within the CD interval. Hence, ProtoSiTex constitutes a new top-performing cluster within interpretable networks.

3) Black-Box Comparison: For the transformer and graph-based baselines (k=11), the Friedman test returned χ_F^2=45.1 and F_F=10.8 ($p < 10^{-6}$), confirming significant global differences. The Nemenyi test (CD=5.87) revealed that RoBERTa-Large and DeBERTaV3 form the leading group with statistically indistinguishable ranks, while both outperformed early transformer variants such as ALBERT and TextGCN. Other transformer models (RoBERTa, ELECTRA, XLNet, BERT, ModernBERT, BertGCN, DistilBERT) lied within the CD range of RoBERTa-Large, indicating no significant pairwise difference.

The statistical analyses jointly establish that ProtoSiTex attains the best overall mean rank among all models, with improvements that are statistically significant relative to previous interpretable frameworks and not significantly different from those of the most powerful black-box transformers. This demonstrates that ProtoSiTex achieves a desirable trade-off between interpretability and predictive accuracy, providing prototype-level transparency without sacrificing statistical performance. The CD diagrams in Fig. D.2 illustrate these findings: PROTOSITEX, RoBERTa-Large, and DeBERTaV3 occupy a shared high-performance cluster, while other models fall significantly behind. Such rank-based robustness across three datasets (HR, IMDb, TweetEval) further validates the stability and generalizability of the proposed architecture.

APPENDIX F
QUALITATIVE ANALYSIS

While we presented a qualitative analysis in the main paper, this section provides a more detailed examination, offering finer-grained insights into the interpretability and effectiveness of ProtoSiTex. Table E.1 show textual representation of learned prototypes along with their corresponding classes/ labels, derived after training on all datasets (HR, IMDb [6], and TweetEVAL [7]). Our analysis reveals that the learned prototypes are well-distributed across each dataset and span all labels. Additionally, we show qualitative results for samples from HR, IMDb, and TweetEVAL in Fig. E.1.

1) Analysis on Hotel Reviews (HR): Fig. E.1(a) demonstrates the interpretability of ProtoSiTex on multi-aspect text. Here, the hotel review referencing honeymoon context directly supports the prototype of "*fostering memorable celebrations and immersive activities*", while review text such as "*hotel has the best view to offer*" closely matches the prototype "*location: scenic and strategically advantageous settings*". ProtoSiTex also provides rich prototype justification for additional aspects and successfully maps the review into *Concierge*, *Personalization*, *Food*, *Location & Surrounding*, *Guest Experience* aspects.

2) Analysis on IMDb: In Fig. E.1(b), ProtoSiTex classifies the input text as *positive*, recognizing its emphasis on how this film version outshines all previous adaptations. The model captures the sentiment of novelty and strength in storytelling, with a strong alignment to relevant prototypes, demonstrating both the interpretability and effectiveness of ProtoSiTex. Likewise, ProtoSiTex predicts *negative* for the sample presented in Fig. E.1(c).

3) Analysis on TweetEval: Fig.s E.1(d)-(g) illustrate the two-way interpretability of ProtoSiTex on TweetEVAL samples. Specifically, Fig. E.1(d), the input text reflects deep emotional pain rooted in past trauma, conveying feelings of depression and grief. This aligns closely with the

TABLE E.1: Textual representative of learned prototypes on All Datasets

Sl.	Prototype Texts	Labels
	HR Dataset (Ours)	
1	Arrival logistics, personalization, and service recovery	Concierge, Guest Experience, Personalization
2	Personalization and navigating external factors	Concerige, Food, Guest Experience, Personalization
3	Inconsistent functionality and safety concerns	Room & Amenities
4	Complimentary room upgrades and enhanced stay benefits	Guest Experience, Personalization, Room & Amenities
5	Exceptional guest experience through proactive service and unexpected delights	Concierge, Guest Experience, Personalization
6	Room quality and prime location versus maintenance and staff issues	Location & Surrounding, Room & Amenities
7	Housekeeping: excellence in service marred by pest issues	Housekeeping
8	Room features, accessibility, and guest privacy concerns.	Room & Amenities
9	Varied impressions of hotel property and service	Guest Experience
10	State of facilities and guest offerings	Room & Amenities
11	Assessing cost-effectiveness and perceived worth	Pricing & Value
12	Well-maintained facilities and friendly service.	Concierge, Guest Experience, Housekeeping, Room & Amenities
13	Exceptional staff and room quality	Concerige, Guest Experience, Housekeeping
14	Location: proximity, environmental noise, and scenic aspects	Location & Surrounding
15	Navigating operational shortcomings and external challenges	Concierge, Guest Experience
16	Cleanliness and quality: rooms, amenities, and breakfast	Food, Housekeeping, Room & Amenities
17	Inconsistent guest experience and room and amenities quality	Guest Experience, Room & Amenities
18	Varied staff performance and property conditions with notable location and dining	Food, Location & Surrounding
19	Variable quality across guest experience and property aspects	Guest Experience
20	Balancing location advantages with pricing concerns	Location & Surrounding, Pricing & Value
21	Overpriced with inadequate service and amenities	Guest Experience, Pricing & Value, Room & Amenities
22	Recurring room and amenity malfunctions and service deficiencies	Housekeeping, Room & Amenities
23	Financial dealings and perceived value	Pricing & Value
24	Inconsistent value for money with unexpected additional costs	Pricing & Value
25	Spacious and family-friendly	Food, Room & Amenities
26	Quality and utility of facilities and in-room provisions	Room & Amenities
27	Varied room types, upgrades, and amenity performance	Room & Amenities
28	Diverse guest insights across hotel operations, location ambiance, and service delivery	Guest Experience, Location & Surrounding
29	Location: scenic and strategically advantageous settings	Location & Surrounding
30	Contrasting guest experience and perceived value	Guest Experience, Pricing & Value
31	Service and billing management	Concierge, Pricing & Value
32	Elevated stays through thoughtful amenities and diverse culinary offerings	Food, Guest Experience
33	Discrepant billing, payment obstacles, and perceived poor value	Concerige, Guest Experience, Pricing & Value
34	Outstanding hospitality, culinary delights, and comfort.	Food, Guest Experience
35	Varying quality in food, facilities, and service	Concerige, Food, Guest Experience
36	Focus on renovation and comfort upgrades	Room & Amenities
37	Handling of booking, billing, and diverse guest requests	Concierge, Personalization, Pricing & Value
38	Property condition, aesthetic appeal, and service quality	Housekeeping, Room & Amenities
39	Prominent on-site dining and accommodation features	Food, Room & Amenities
40	Diverse dining experiences and efficient guest services	Concierge, Food
41	Fostering memorable celebrations and immersive activities	Guest Experience, Personalization
42	Mixed perceptions of room condition, amenities, and service quality relative to cost	Pricing & Value, Room & Amenities
43	Location: a decisive factor for guests	Location & Surrounding
44	Hotel's location and on-site environment: accessibility, amenities, and challenges	Location & Surrounding, Room & Amenities
45	Food & beverage: value, variety, and pricing transparency	Food, Pricing & Value
46	Staff engagement and support	Concierge
47	Personalized attention and special arrangements	Guest Experience, Personalization
48	Modern amenities and varied room conditions with notable maintenance and service lapses	Housekeeping, Room & Amenities
	IMDB Dataset [6]	
1	Unveiling the depth of human emotion through diverse narratives	Positive
2	Discerning appraisals of cinematic,theatrical performances, adaptations	Positive
3	Celebrating unconventional and engrossing storytelling	Positive
4	Potent narratives of human struggle and unflinching reality	Positive
5	Diverse perspectives and defended convictions	Positive
6	Cinema of intense debate and baffling choices	Positive
7	Memorable, albeit flawed, cinematic experiences	Negative
8	Films with commendable elements, yet falling short of their potential	Negative
	TweetEVAL Dataset [7]	
1	Irritation and frustration	Anger
2	Anger and disgust	Anger
3	Irritation and disrespect triggered outbursts	Anger
4	Humorous and offended outbursts	Anger
5	Contentment and reflection	Joy
6	Cuddling and positivity promote well-being	Joy
7	Gratitude and joyful appreciation	Joy
8	Regret, resilience, and hope amidst adversity	Optimism
9	Minor discomforts and empathetic support	Optimism
10	Gloomy discomfort	Sadness
11	Disappointment and discouragement	Sadness
12	Existential dread and loss	Sadness

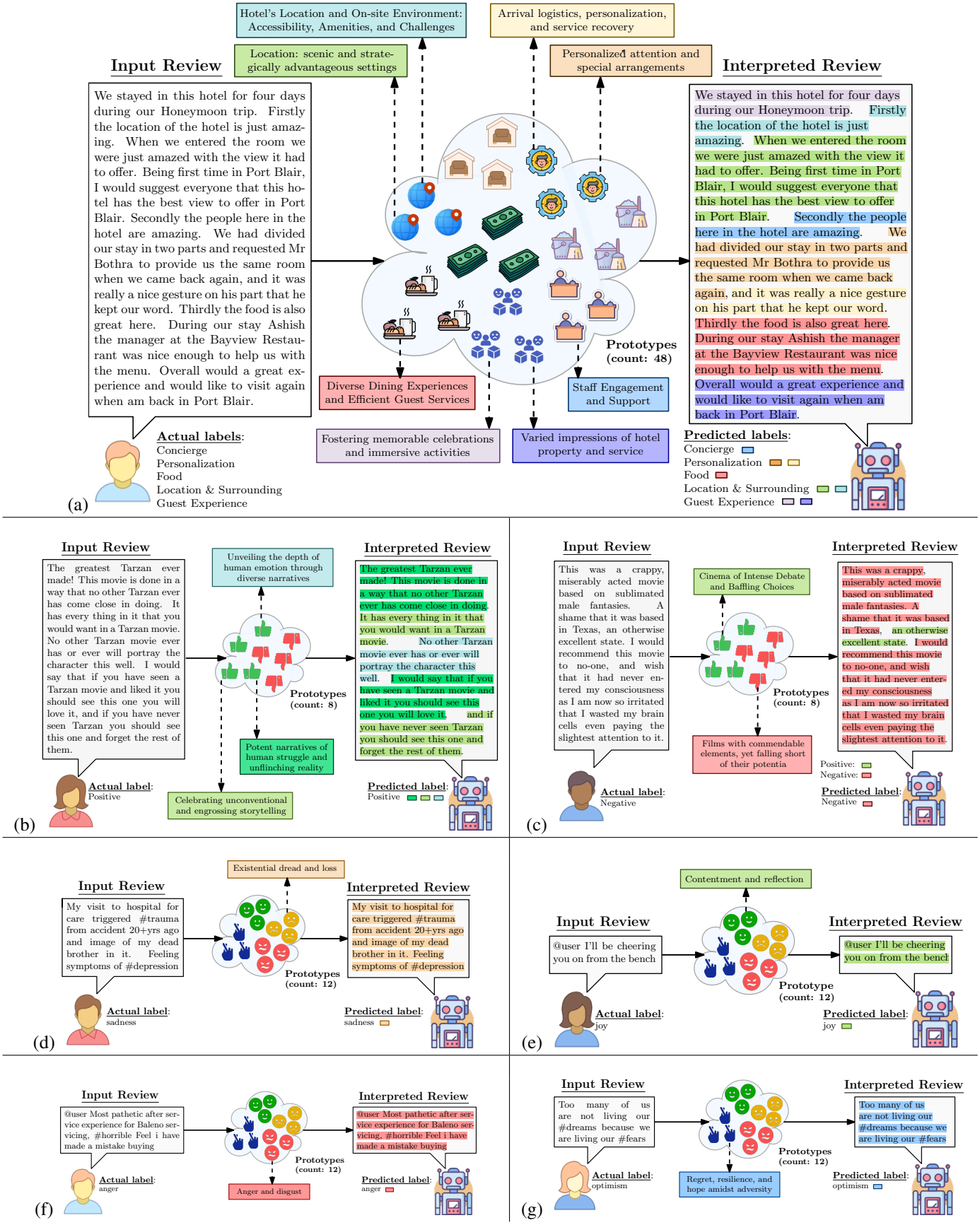


Fig. E.1: Qualitative analysis of ProtoSiTex on samples from (a) HR, (b)-(c) IMDb [6], and (d)-(g) TweetEVAL [7]. Best viewed in color.

prototype capturing themes of "*existential dread and loss*", showcasing ProtoSiTex's ability to contextualize emotional content through semantically coherent prototypes.

REFERENCES

[1] M. Nauta *et al.*, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," *ACM CSUR*, vol. 55, no. 13s, pp. 1–42, 2023.

[2] H. Monke *et al.*, "From confusion to clarity: ProtoScore-a framework for evaluating prototype-based xai," in *ACM FAccT*, 2025, pp. 2215–2231.

[3] S. Kornblith *et al.*, "Similarity of neural network representations revisited," in *ICML*. PMlR, 2019, pp. 3519–3529.

[4] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *JMLR*, vol. 7, no. Jan, pp. 1–30, 2006.

[5] O. Rainio and Others, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, no. 1, p. 6086, 2024.

[6] A. L. Maas *et al.*, "Learning word vectors for sentiment analysis," in *ACL*, 2011, pp. 142–150.

[7] F. Barbieri *et al.*, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *EMNLP*, 2020, pp. 1644–1650.