LGSA: Label Geometry Structuring and Aligning for Hierarchical Text Classification

Anonymous ACL submission

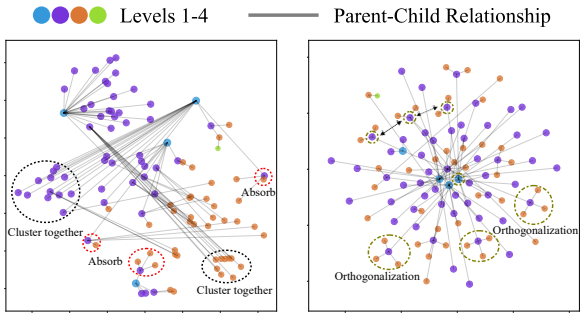
Abstract

Existing hierarchical text classification (HTC) methods typically use prompt tuning or contrastive learning to inject the label hierarchy into a model as prior knowledge to implicitly learn label embeddings for classification. However, such implicit learning fails to accurately reflect label geometry (i.e., feature spatial distribution of label embeddings), as it does not model hierarchy-aware geometric relations among labels. To address this issue, we propose a novel two-stage label geometry structuring and aligning framework, termed LGSA, which transforms the label hierarchy from an implicit prior into an explicit embedding. First, we propose a hierarchical geometric structuring (HGS) module that leverages a general orthogonal frame (GOF) to reconstruct an explicit label geometry conforming to the label hierarchy. The label geometry is then treated as a label prototype to guide model training. To facilitate the guidance, we thereby propose a hierarchical geometric aligning (HGA) module as a regularization term to align label geometry learned by the model with the explicit label prototype. Experiments on three real-world HTC datasets confirm that LGSA consistently outperforms existing state-of-the-art methods. The code and models are available at https://anonymous.4open.science/r/LGSA-1E0C.

1 Introduction

Hierarchical text classification (HTC), an important task in assigning texts to a predefined label hierarchy (Sun et al., 2004), has wide-ranging applications in domains such as academic paper classification (Kowsari et al., 2017) and product categorization (Cevahir and Murakami, 2016). However, real-world taxonomies are usually deep and large-scale, with complex parent-child and sibling relations among labels (Mao et al., 2019), which makes accurate HTC still an open challenge.

To tackle this challenge, researchers have explored various approaches to model the label hierarchy. Prompt tuning-based methods (Wang et al., 2022b; Xiong et al., 2024) construct hierarchy-aware textual templates to reframe classification as a cloze-style task, which inject the label hierarchy as prior knowledge into pre-trained language models (PLMs). Contrastive learning-based methods (Wang et al., 2022a; Yu et al., 2023) construct positive and negative pairs at the instance or label level, and then use the label hierarchy to determine sample similarities, allowing learning of label embeddings and hierarchical relations in the feature space.



(a) Existing Methods (b) Our Method

Figure 1: Feature spatial distribution visualization of labels learned by (a) existing methods and (b) our proposed method. Levels 1 to 4 represent the hierarchy of labels, with Level 1 at the top and Level 4 at the leaf. The existing methods exhibit a distorted geometric structure, while ours generates labels with a geometric structure that follows the label hierarchy, with clear geometric separation.

However, such approaches implicitly learn label embeddings, ignoring hierarchy-aware geometric relations among labels when modeling complex label hierarchies in HTC. The resulting representation fails to accurately reflect the feature spatial distribution of label embeddings (which we refer to as label geometry), and may even lead to distortions of the label geometry, especially for long-tailed label distributions. The distortions are reflected in the

following two aspects: 1) **Structural Confusion**: The learned feature spatial distribution fails to create a topology consistent with the label hierarchy. Semantic-related labels (especially fine-grained sibling labels) are often mapped to close positions in the feature space, making them hard to distinguish (as shown in Figure 1(a), where level 3 labels are clustered together). 2) **Representation Collapse**: Tail labels with limited samples often lead to indistinct embedding representations. During optimization, their representations are easily absorbed by corresponding parent or head labels, leading to a decrease in the discriminability of tail labels. Although recent works have attempted to improve the feature spatial distribution by adding additional constraints, such as learnable decision boundaries (Kim et al., 2024) or structural entropy (Zhu et al., 2024; Liu et al., 2025), these constraints are either local and highly data-dependent or global but agnostic to hierarchical structure. Both constraints are based on implicitly learned label embeddings, failing to provide a hierarchy-aware geometric structure and thus can not fundamentally resolve the issues of structural confusion and representation collapse.

To address the above challenges, we propose a novel two-stage framework, called label geometry structuring and aligning (LGSA) for the HTC task. It transforms the label hierarchy from a passively learned implicit prior into a structurally clear and explicit embedding. During model initialization, we propose a hierarchical geometric structuring (HGS) module. Specifically, HGS first generates initial semantic embeddings for each label. It then reconstructs an explicit label geometry that conforms to the label hierarchy in a top-down manner by leveraging the general orthogonal frame (GOF) (Dang et al., 2023) and the Gram-Schmidt orthogonalization process (Golub and Van Loan, 2013), providing a stable and structurally meaningful geometric position in feature space for each label. The reconstructed label geometry is used as a label prototype to guide subsequent model training. During model training, we propose a hierarchical geometric aligning (HGA) module to guide the model in learning the predefined label geometry. Unlike methods that optimize only global spectral properties (e.g., singular spectrum smoothing (Liu et al., 2025)), HGA enforces explicit structural alignment, ensuring that learned embeddings strictly adhere to the directional relationships of the hierarchy rather than merely smoothing feature discrepancies. Specifically, HGA serves as a geometric regularizer for label embeddings by imposing complementary geometric constraints that align the learned label geometry with the explicit prototype. In practice, a global geometric constraint ensures geometric orthogonality among head labels, while a set of local geometric constraints continuously calibrates specific hierarchical topological relations (e.g., parent-child and sibling relationships), thereby avoiding structural confusion and representation collapse.

Our contributions are summarized as follows:

- We propose a novel two-stage framework LGSA, to achieve accurate HTC in a complex label hierarchy.
- We propose a HGS module to construct a prior label prototype with an explicit topological structure of the label hierarchy.
- We design a HGA module as a geometric regularizer to align label embeddings with the prior label prototype.
- Extensive experiments on three real-world HTC datasets demonstrate the effectiveness of the proposed LGSA.

2 Related Work

2.1 Hierarchical Text Classification

Research in HTC has primarily focused on leveraging label structural dependencies to guide semantic representation learning. Early works employ reinforcement learning (Mao et al., 2019) or capsule networks (Aly et al., 2019) to model label relationships or decision paths for top-down text classification. Subsequent hierarchy-aware approaches, such as the dual-encoder frameworks proposed by Zhou et al. (2020) and Deng et al. (2021), separately encode the text and the label structure to facilitate their interaction for final label prediction. However, these methods often fail to capture complex semantic-structural interactions and result in low-quality label embeddings in imbalanced data settings. More recent work has been devoted to the fine-tuning paradigm. Prompt-tuning methods integrate hierarchical information into text templates to reframe HTC as a masked language modeling task, utilizing techniques such as dynamic virtual templates (Wang et al., 2022b), dual prompts (Xiong et al., 2024), or local hierarchy integration (Kong

et al., 2025). Concurrently, contrastive learning approaches (Wang et al., 2022a; Ji et al., 2023) have been employed to strengthen semantic representations and implicitly structure the feature space. Despite their advances, the feature space learned by these methods is highly dependent on the data distribution. Consequently, the issues of structural confusion and representation collapse persist in these methods. In this paper, we decouple label embeddings from the data distribution in an explicit manner, avoiding passive feature space construction.

2.2 Geometric Structure

The geometric structure of sample or label embeddings in the feature space is a central topic in representation learning. Papyan et al. (2020) first revealed that under balanced data conditions, the classifier weights spontaneously converge to an ideal simplex equiangular tight frame (ETF). This theory has been later extended by Dang et al. (2023, 2024) to imbalanced settings, where classifier weights are demonstrated to converge to a distorted GOF, wherein the norms of tail labels are severely shortened, providing a geometric explanation for the representation collapse. To rectify such geometric distortion, some studies have utilized the ideal ETF or a balanced GOF as a strong geometric prior to actively guide the model in forming a more balanced and highly discriminative feature space during training. For instance, Gao et al. (2024) enforced an alignment between representations learned from imbalanced data and a target ETF, while Pham et al. (2025) employed the GOF to ensure geometric separation between labels. However, these methods assume that labels are independent and thus are only applicable to flat classification, failing to address the complex hierarchical dependencies inherent in HTC. To our knowledge, our work is the first to design and construct a label geometry explicitly tailored for the label hierarchy of HTC.

3 Methodology

3.1 Problem Definition

In HTC, the set of labels $\mathcal{Y}$ is organized into a hierarchy, formally defined as a tree-structured directed acyclic graph (DAG) $\mathcal{H} = (\mathcal{Y}, \mathcal{E})$, where each node represents a label and the edges in $\mathcal{E}$ represent parent-child relationships. Given a text X, the task is to predict a set of labels $Y \subseteq \mathcal{Y}$. The prediction must be hierarchically consistent, meaning that if the i-th label $y_i \in \mathcal{Y}$ is predicted, all of its ancestor labels must also be in Y. This constraint ensures that the final label set corresponds to one or more valid root-to-node paths in $\mathcal{H}$.

3.2 Overall Architecture

In this section, we will detail our proposed two-stage framework, LGSA. Its overall architecture is illustrated in Figure 2, which first constructs a prior geometric prototype in the feature space via HGS (Section 3.3). Subsequently, it aligns the learned label embeddings with the prior geometric prototype using the HGA module (Section 3.4).

3.3 Hierarchical Geometric Structuring

To resolve structural confusion and representation collapse, label geometry necessitates three critical conditions: (1) head labels require global orthogonality to form a clear geometric separation, laying the foundation for distinct hierarchical branches; (2) sibling labels require orthogonality in their unique semantic features to distinguish them from one another under the same parent label; (3) parent-child labels demand directional consistency in the feature space to accurately reflect hierarchical relationships. Based on these conditions, we propose the HGS module to construct a prior label prototype of labels during the model initialization stage. The geometric construction consists of the following steps:

Step1: Label Initialization. Following previous work (Wang et al., 2022b; Liu et al., 2025), we first utilize the word embedding layer of BERT (Devlin et al., 2019) to generate semantic embeddings e'_i for each label. Next, we integrate the label hierarchy into the semantic embeddings via a graph attention network (GAT) (Veličković et al., 2018) to obtain the initial label embeddings e_i. The label initialization operation is defined as follows:

$$e'_i = f_{BERT\_embed}(y_i) \quad (1)$$

$$\{e_i\}_{i=1}^{|\mathcal{Y}|} = f_{GAT}(\{e'_i\}_{i=1}^{|\mathcal{Y}|}, A) \quad (2)$$

where $f_{BERT\_embed}(\cdot)$ and $f_{GAT}(\cdot)$ represent the word embedding operation of BERT and the graph encoding operation of GAT, respectively; and A represents the adjacency matrix of the labels.

Step 2: Global Orthogonalization. While the rich semantic information captured in Step 1 is indispensable, distinct geometric separation among

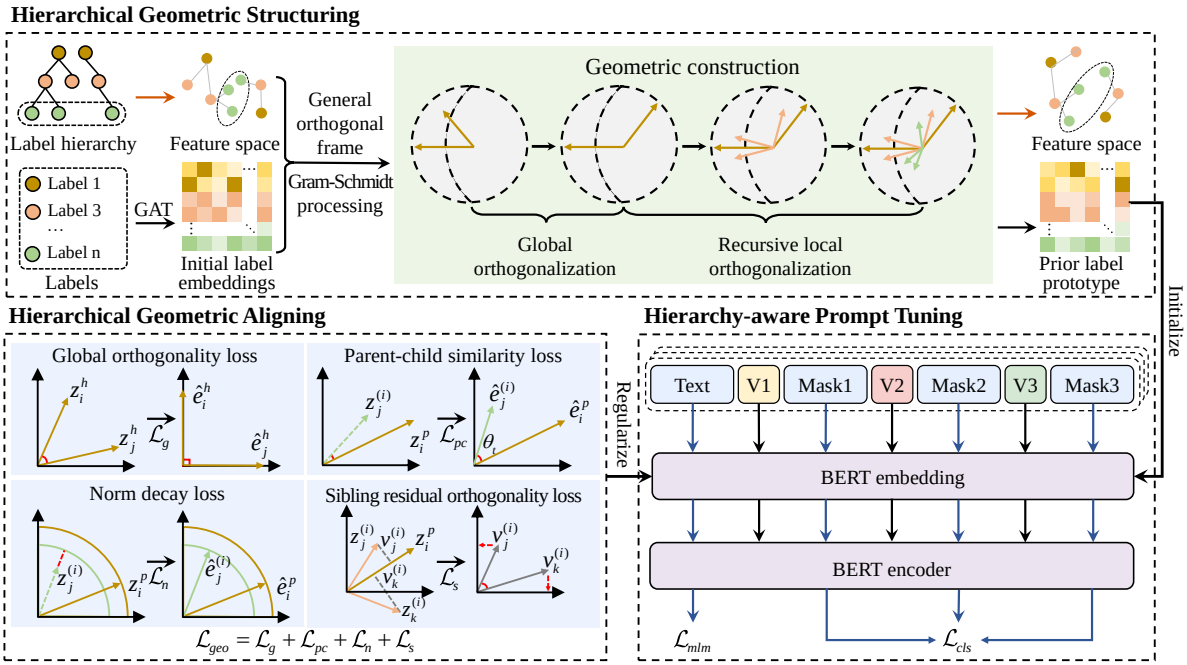


Figure 2: Overall architecture of the proposed LGSA. It consists of two innovative modules: (1) HGS, which constructs an explicit label prototype via orthogonalization as a structural prior; and (2) HGA, which strictly aligns learned embeddings with the prototype to mitigate structural confusion and representation collapse.

head labels is equally critical for defining clear hierarchical branches. Thus, we orthogonalize the embeddings of the head labels to enhance their independence in the feature space and to establish clear boundaries for the main branches of the hierarchy. Specifically, given the head label set Φ_h and their initial label embeddings $\{e_i^h\}_{y_i \in \Phi_h}$, we first apply the Gram-Schmidt orthogonalization method to obtain orthonormal bases $\{b_i^h\}_{y_i \in \Phi_h}$. To preserve the semantic strength represented by the norm of the original embeddings, we then rescale these bases back to the original norms. This process is formulated as follows:

$$\begin{aligned} b_i^h &= \mathrm{GramSchmidt}(e_i^h) \\ &= e_i^h - \sum_{j=1}^{h-1} \frac{e_i^h \cdot b_j^h}{\|b_j^h\|_2^2} b_j^h \end{aligned} \tag{3}$$

$$\hat{e}_i^h = \|e_i^h\|_2 b_i^h \tag{4}$$

where $\hat{e}_i^h$ denotes the label prototype of the i-th head label after orthogonalization and norm rescaling; $\|\cdot\|_2$ denotes the Euclidean (L2) norm.

Step 3: Recursive Local Orthogonalization. Following the global orthogonalization of head labels, we propagate the process top-down through the label hierarchy, maintaining semantic consistency between parent-child labels while disentangling sibling labels. For the i-th parent label, we denote the embedding of its j-th child label by $e_j^{(i)}$ and decompose it into two semantically distinct components: a projection embedding $p_j^{(i)}$ along the parent label direction, representing the inherited common semantics, and the unique semantic features $r_j^{(i)}$, which are referred to as the residual embedding. We then apply the Gram-Schmidt orthogonalization method (as used in Step 2) to $r_j^{(i)}$ to ensure that the sibling labels remain orthogonal within the subspace defined by the parent label. To prevent semantic loss, the norms of the original residual embeddings are preserved during construction, allowing each child label to inherit the semantics of its parent label while maintaining an independent geometric representation, calculated as:

$$p_j^{(i)} = \mathrm{proj}(e_j^{(i)}) = \frac{\hat{e}_i^p \cdot e_j^{(i)}}{\|\hat{e}_i^p\|_2^2} \hat{e}_i^p \tag{5}$$

$$r_j^{(i)} = e_j^{(i)} - p_j^{(i)} \tag{6}$$

$$\hat{e}_j^{(i)} = \mathrm{GramSchmidt}(r_j^{(i)}) + p_j^{(i)} \tag{7}$$

where $\hat{e}_i^p$ and $\hat{e}_j^{(i)}$ denote the label prototypes of i-th parent label and its j-th child label obtained after orthogonalization

3.4 Hierarchical Geometric Aligning

To explicitly incorporate the constructed label prototype into the representation learning process, we

propose the HGA module. Specifically, it employs a hierarchical geometric regularization loss $\mathcal{L}_{geo}$, which comprises four specific losses as geometric constraints to collectively regularize the label embeddings. Crucially, these constraints are organized to address distinct topological scopes in a complementary manner. The global orthogonality loss $\mathcal{L}_g$ enforces orthogonality among head labels, enabling geometric separation and providing independent subspaces for local geometric constraints. Subsequently, local geometric constraints recursively calibrate specific hierarchical topological relations. Specifically, the parent-child similarity loss $\mathcal{L}_{pc}$ and the norm decay loss $\mathcal{L}_n$ jointly preserve vertical consistency (in terms of direction and magnitude), while the sibling residual orthogonality loss $\mathcal{L}_s$ guarantees horizontal discriminability among sibling labels. Thus, the $\mathcal{L}_{geo}$ is the sum of the four losses above, formally expressed as:

$$\mathcal{L}_{geo} = \mathcal{L}_g + \mathcal{L}_{pc} + \mathcal{L}_n + \mathcal{L}_s \tag{8}$$

Global Orthogonality Loss $\mathcal{L}_g$. To guarantee the geometric separation of head labels, this loss enforces explicit orthogonality by minimizing the pairwise cosine similarity between their embeddings. This ensures that the major branches of the label hierarchy remain mathematically independent, formulated as:

$$\mathcal{L}_g = \frac{2}{N_h} \sum_{y_i, y_j \in \Phi_h, i<j} \left(\frac{z_i^h \cdot z_j^h}{\|z_i^h\|_2 \|z_j^h\|_2 + \epsilon} \right)^2 \tag{9}$$

where ϵ is a very small constant to prevent the denominator from becoming zero; z_i^h and z_j^h denote the learned embeddings of the i-th and j-th head labels during the training process, respectively; and $N_h = |\Phi_h|(|\Phi_h| - 1)$.

Parent-Child Similarity Loss $\mathcal{L}_{pc}$. The "is-a" hierarchical relation geometrically requires directional consistency to capture intrinsic semantic dependencies. To explicitly model this inheritance during training, this loss constrains child label embeddings to lie within a narrow cone centered on their parent label embedding. This spatial constraint establishes a geometric tolerance margin for semantic variation. Specifically, we apply a squared penalty to deviations from this margin to ensure that child label embeddings closely follow the direction of their parent label embedding, thus geometrically modeling the inheritance relationship between parent-child labels:

$$\mathcal{L}_{pc} = \frac{1}{|\mathcal{P}|} \sum_{(y_i, y_j) \in \mathcal{P}} \left(\frac{z_j^c \cdot z_i^p}{\|z_j^c\|_2 \|z_i^p\|_2 + \epsilon} - \cos(\theta_t) \right)^2 \tag{10}$$

where $\mathcal{P}$ is the set of all parent-child label pairs; and θ_t denotes the target angle serving as a tolerance margin.

Norm Decay Loss $\mathcal{L}_n$. To prevent the embeddings of tail labels from losing discriminability due to small norms caused by imbalanced data, this loss introduces a norm decay constraint between parent and child labels. Specifically, it ensures that the norm of a child label maintains a reasonable ratio relative to that of its parent, preventing semantic degradation (i.e., vanishing norms):

$$\mathcal{L}_n = \frac{1}{|\mathcal{P}|} \sum_{(y_i, y_j) \in \mathcal{P}} \mathrm{ReLU}(\gamma \|z_i^p\|_2 - \|z_j^{(i)}\|_2) \tag{11}$$

where γ is a norm decay factor. The ReLU function (Nair and Hinton, 2010) ensures that the penalty is applied only when the norm of the child label is smaller than the rescaled norm of the parent label.

Sibling Residual Orthogonality Loss $\mathcal{L}_s$. While sibling labels share a parent label, their effective distinction relies on clear geometric separation. To mitigate the confusion caused by the dense clustering of tail labels, this loss enforces orthogonality among their residual embeddings, ensuring that they have independent and distinguishable features:

$$v_j^{(i)} = z_j^{(i)} - \mathrm{proj}(z_j^{(i)}) \tag{12}$$

$$\mathcal{L}_s = \frac{1}{|\mathcal{F}|} \sum_{y_i \in \mathcal{F}} \left[\frac{2}{|\mathcal{C}_i|(|\mathcal{C}_i| - 1)} \sum_{y_j, y_k \in \mathcal{C}_i, j<k} \left| \frac{v_k^{(i)} \cdot v_j^{(i)}}{\|v_k^{(i)}\|_2 \|v_j^{(i)}\|_2 + \epsilon} \right| \right] \tag{13}$$

where $\mathcal{F}$ is the set of parent labels with at least two child labels; $\mathcal{C}_i$ denotes the child label set of the i-th parent label.

3.5 Multi-task Training

Multitask training is a weighted combination of classification loss ($\mathcal{L}_{cls}$), masked language modeling loss ($\mathcal{L}_{mlm}$), and hierarchical geometric alignment loss ($\mathcal{L}_{geo}$). The final objective loss function to be minimized can be defined as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{mlm} + \lambda \mathcal{L}_{geo} \tag{14}$$

where λ represents the hyperparameter that controls the weight of the $\mathcal{L}_{geo}$.

Model	WOS		RCV1-V2		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
HiAGM (Zhou et al., 2020)	85.82	80.28	83.96	63.35	74.97	60.83
HTCInfoMax (Deng et al., 2021)	85.58	80.05	83.51	62.71	-	-
HiMatch (Chen et al., 2021)	86.20	80.53	84.73	64.11	-	-
HGCLR (Wang et al., 2022a)	87.11	81.20	86.49	68.31	78.86	67.96
HJCL (Yu et al., 2023)	-	-	87.04	70.49	80.52	70.02
HILL (Zhu et al., 2024)	87.28	81.77	87.31	70.12	80.47	69.96
HiSR (Zhou et al., 2025)	87.52	82.04	87.59	70.72	80.32	70.11
HPT (Wang et al., 2022b)	87.16	81.93	87.26	69.53	80.42	70.42
DPT (Xiong et al., 2024)	87.25	81.51	87.76	70.78	80.56	70.28
COPHTC (Cai et al., 2024b)	87.42	82.09	87.32	70.12	80.77	70.59
NERHTC (Cai et al., 2024a)	87.42	81.93	87.50	69.76	80.97	70.99
SIHTC (Liu et al., 2025)	87.65	82.02	87.83	70.36	80.84	70.87
LGSA (proposed)	**87.79**	**82.47**	**87.90**	**71.35**	**81.19**	**71.35**

Table 1: Experimental results on three HTC datasets. In each column, the best value is in bold, the second best value is underlined.

4 Experiments

4.1 Experiment Setup

Datasets and Evaluation Metrics. We conduct experiments on three datasets: Web-of-Science (WOS) (Kowsari et al., 2017), RCV1-V2 (Lewis et al., 2004), and NYT (Sandhaus, 2008) to evaluate the effectiveness of our proposed LGSA. To ensure a fair comparison, we adopt a data pre-processing technique described in Chen et al. (2021) and Wang et al. (2022b), with data cleaning and splitting into training, validation, and test sets. According to the established evaluation practices (Gopal and Yang, 2013), two standard evaluation metrics, Macro-F1 and Micro-F1 (Yang, 1999), are chosen to assess the performance of the model. *Dataset statistics are provided in Appendix A.*

Implementation Details. Building upon previous work (Liu et al., 2025), we adopt the HPT model (Wang et al., 2022b) as the base model of the proposed LGSA framework, which is trained with a batch size of 16, using an Adam optimizer with a learning rate of 1e-4. To enable stable geometric regularization, we employ a warmup strategy for the hyperparameter λ, which linearly increases from zero to the preset value during the first 10% of training steps. *Further details regarding the base model and comprehensive hyperparameter settings are provided in Appendix B.*

Baselines. We compare LGSA with the following advanced baselines: (1) **Hierarchy-aware Models.** HiAGM (Zhou et al., 2020), HTCInfoMax (Deng et al., 2021), and HiMatch (Chen et al., 2021) utilized dual-encoder architectures to model the interaction between text and label hierarchies. (2) **Contrastive Learning-based Models.** HGCLR (Wang et al., 2022a), HJCL (Yu et al., 2023), HILL (Zhu et al., 2024), and HiSR (Zhou et al., 2025) employed contrastive learning to enhance feature representations by incorporating hierarchical information. (3) **Prompt Tuning-based Models.** HPT (Wang et al., 2022b) pioneered the application of prompt tuning in HTC. Building on this line of work, subsequent works extended HPT along several directions: DPT (Xiong et al., 2024) proposed dual-prompt mechanisms, COPHTC (Cai et al., 2024b) incorporated contrastive constraints, NERHTC (Cai et al., 2024a) integrated named entity recognition, and SIHTC (Liu et al., 2025) exploited structural entropy.

4.2 Main Results

The main experimental results are summarized in Table 1. Among all compared baselines, LGSA achieves the best performance on the three HTC datasets while introducing only a small number of additional parameters over the base model HPT.

On the WOS dataset with a label depth of 2, LGSA outperforms the strongest baseline, achieving absolute gains of 0.14% in Micro-F1 over SIHTC and 0.38% in Macro-F1 over COPHTC. On the RCV1-V2 dataset with a label depth of 4, LGSA demonstrates significant advantages, particularly in handling imbalanced data. It outperforms HPT by 0.64% in Micro-F1 and a substantial 1.82% in Macro-F1. Furthermore, LGSA surpasses the previous best-performing baseline, DPT, by 0.57%

Module Variants	RCV1-V2		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
LGSA (proposed)	**87.90**	**71.35**	**81.19**	**71.35**
-r.m. HGS & HGA	87.51	69.08	80.45	70.00
-r.m. HGS	87.70	70.20	80.56	70.43
-r.m. HGA	87.51	70.71	80.83	70.80

Table 2: Ablation study of key components in LGSA on RCV1-V2 and NYT datasets.

Module Variants	RCV1-V2		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
LGSA (proposed)	**87.90**	**71.35**	**81.19**	**71.35**
-r.m. Step 2	87.84	70.68	80.84	70.96
-r.m. Step 3	87.32	70.11	81.02	70.75
-r.p. Random	87.32	69.40	79.63	68.55
-r.p. Frozen	87.84	71.01	80.93	70.57

Table 3: Ablation study of geometric construction strategies within HGS. "Random" indicates that the label embeddings are initialized from the standard normal distribution. "Frozen" indicates the label prototypes are not updated during the training process.

Loss Variants	RCV1-V2		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1
LGSA (proposed)	**87.90**	**71.35**	**81.19**	**71.35**
-r.m. $\mathcal{L}_g$	87.63	70.87	80.84	70.56
-r.m. $\mathcal{L}_{pc}$	87.83	70.86	81.04	71.04
-r.m. $\mathcal{L}_n$	87.85	70.97	81.01	71.17
-r.m. $\mathcal{L}_s$	87.85	70.82	80.69	70.61

Table 4: Ablation study of four geometric constraints within HGA.

in Macro-F1. The pronounced improvement in Macro-F1 strongly suggests that LGSA effectively mitigates the representation collapse of tail labels. On the NYT dataset with a label depth of 8, LGSA demonstrates superior robustness in modeling complex dependencies. It consistently outperforms all baselines. Specifically, it surpasses the strongest baseline, NERHTC, by 0.22% and 0.36% in Micro-F1 and Macro-F1, respectively. This indicates that the proposed explicit geometric modeling approach is effective for alleviating structural confusion even in deep hierarchies.

4.3 Ablation Study

Key Modules in LGSA. The ablation results for key modules are presented in Table 2. Removing both HGS and HGA yields the largest performance drop, reducing Macro-F1 by 2.27% on RCV1-V2 and 1.35% on NYT, indicating that without explicit geometric modeling the model collapses into a generic classifier that fails to capture complex hierarchical dependencies. Removing HGS alone causes a clear degradation in Macro-F1 (1.15% and 0.92%), suggesting that HGS produces high-quality label prototypes that facilitate the subsequent HGA module. Likewise, ablating HGA results in marked declines on both datasets, highlighting the importance of continuous geometric alignment during optimization. Together, these results validate that HGS and HGA are mutually reinforcing.

Geometric Construction in HGS. Table 3 validates the HGS design strategies. First, the Random variant suffers the largest drop in Macro-F1 (1.95% on RCV1-V2 and 2.80% on NYT), confirming that geometric structuring is ineffective without the semantic embeddings provided by the Step 1. Second, removing Step 3 impacts performance in Macro-F1 more severely than removing Step 2 on both datasets (e.g., a drop of 1.24% vs. 0.67% on RCV1-V2). This suggests that while global orthogonalization ensures distinct geometric separation among head labels, disentangling fine-grained sibling dependencies is the more critical bottleneck in both imbalanced and deep hierarchies. Finally, the learnable HGS outperforms the frozen variant, indicating that while the label prototype provides a robust foundation, fine-tuning label embeddings during training is essential for optimal alignment.

Geometric Constraint in HGA. Table 4 details the contributions of the four geometric constraints within HGA. On NYT dataset, removing the global constraint $\mathcal{L}_g$ results in the largest performance drop (0.79% in Macro-F1). This validates that enforcing global orthogonality among head labels is critical for providing independent subspaces to alleviate structural confusion. On RCV1-V2 dataset, removing the sibling loss $\mathcal{L}_s$ leads to a marked drop (0.53% in Macro-F1), confirming its role in ensuring horizontal discriminability among fine-grained sibling labels. In contrast, removing $\mathcal{L}_{pc}$ or $\mathcal{L}_n$ yields more moderate declines. These findings suggest that establishing clear geometric boundaries through $\mathcal{L}_g$ and $\mathcal{L}_s$ is the primary driver of performance, whereas $\mathcal{L}_{pc}$ and $\mathcal{L}_n$ serve as auxiliary terms to preserve vertical hierarchical consistency.

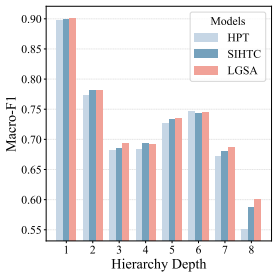
4.4 Exploratory Study

4.4.1 Research on Imbalanced Hierarchy

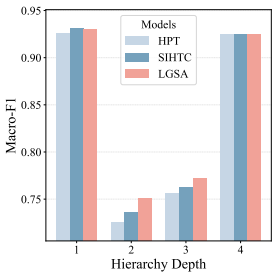
To assess the robustness of LGSA against structural complexity and data imbalance, we analyze its performance across varying hierarchy depths

Metric	WOS			RCV1-V2			NYT		
	HPT	SIHTC	LGSA	HPT	SIHTC	LGSA	HPT	SIHTC	LGSA
Train Time (min)	11.35	12.10	11.62	8.33	9.18	8.20	8.98	10.18	9.47
Eval Time (min)	2.72	4.50	2.70	1.00	1.68	0.95	5.42	9.92	5.45
Params (M)	114.94	114.94	115.05	114.92	114.92	115.00	114.97	114.97	115.10
Memory (M)	20992	21299	20992	20890	22835	20890	21094	21606	21094

Table 5: Comparison of computational efficiency across three datasets. LGSA achieves comparable or better efficiency than the base model HPT, while significantly outperforming SIHTC in both speed and memory usage.

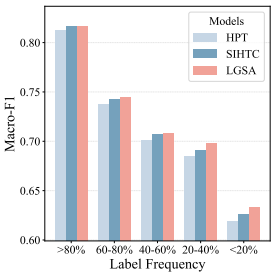


(a) NYT

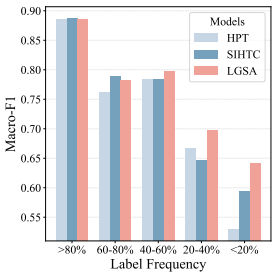


(b) RCV1-V2

Figure 3: Performance comparison of models across different hierarchy depths.



(a) NYT



(b) RCV1-V2

Figure 4: Performance comparison on the long-tailed distribution. Labels are grouped into five intervals based on their label frequency. >80% denotes the top 20% most frequent labels, and the same rule applies to other intervals.

and label frequencies, following the experimental setup outlined in Wang et al. (2022b). As illustrated in Figures 3(a) and 3(b), LGSA consistently outperforms or remains competitive with HPT and SIHTC across all hierarchy depths on both NYT and RCV1-V2 datasets, demonstrating its superior stability in deep structures. Furthermore, grouping labels by label frequency (Figure 4(a) and 4(b)) reveals that LGSA achieves significant gains, particularly on data-scarce tail labels, where baseline models typically struggle. These results confirm that imposing explicit geometric priors and constraints effectively alleviates long-tailed effects, fostering more generalizable and discriminative representations for tail labels even within complex label hierarchies.

4.4.2 Computational Efficiency

We evaluate the computational efficiency of LGSA against the base model HPT and the strong baseline SIHTC on a single NVIDIA RTX 3090 (batch size 16). As shown in Table 5, LGSA demonstrates superior parameter efficiency, introducing negligible parameter overhead compared to HPT and confirming the efficiency of our geometric modules. In particular, LGSA maintains a memory footprint identical to HPT, avoiding the substantial overhead of SIHTC (e.g., 2000MB increase on RCV1-V2). Furthermore, LGSA matches HPT's training speed and outperforms SIHTC in evaluation time. These findings confirm that LGSA effectively balances high performance with computational cost, achieving substantial performance gains without imposing extra demands on hardware resources. *More experimental results are listed in the Appendix C.*

5 Conclusion

In this paper, we propose a novel two-stage framework, LGSA, to address the persistent issues of structural confusion and representation collapse in HTC. Unlike existing methods that implicitly learn label embeddings, LGSA decouples label geometry from data distribution. In the model initialization stage, by introducing the GOF, HGS reconstructs the label embeddings into a stable geometric structure that reflects the label hierarchy. In the model training stage, LGSA uses the HGA module as a geometric regularization term to align the label embeddings with the predefined geometric structure, ensuring that the model maintains discriminability even for long-tailed distributions. Extensive experiments show that LGSA achieves state-of-the-art performance on three real-world HTC datasets, demonstrating that actively constructing a geometric structure is an effective paradigm.

Limitation

In our work, LGSA successfully incorporates a predefined and stable orthogonal label prototype to guide representation learning for labels, effectively alleviating structural confusion and representation collapse. While this rigid structure ensures distinctiveness, it may not fully capture the nuanced semantic affinities, such as varying distances among sibling labels. In future work, we plan to explore dynamic geometric structuring methods that can adaptively learn the optimal topology during training, thereby eliminating the need for rigid pre-alignment.

References

Rami Aly, Steffen Remus, and Chris Biemann. 2019. Hierarchical multi-label classification of text with capsule networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 323–330, Florence, Italy. Association for Computational Linguistics.

Fuhan Cai, Duo Liu, Zhongqiang Zhang, Ge Liu, Xiaozhe Yang, and Xiangzhong Fang. 2024a. NER-guided comprehensive hierarchy-aware prompt tuning for hierarchical text classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 12117–12126, Torino, Italy. ELRA and ICCL.

Fuhan Cai, Zhongqiang Zhang, Duo Liu, and Xiangzhong Fang. 2024b. COPHTC: Contrastive learning with prompt tuning for hierarchical text classification. In *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5400–5404, Seoul, Korea. IEEE.

Ali Cevahir and Koji Murakami. 2016. Large-scale multi-class and hierarchical product categorization for an E-commerce giant. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 525–535, Osaka, Japan. The COLING 2016 Organizing Committee.

Haibin Chen, Qianli Ma, Zhenxi Lin, and Jiangyue Yan. 2021. Hierarchy-aware label semantics matching network for hierarchical text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4370–4379, Online. Association for Computational Linguistics.

Hien Dang, Tho Tran, Tan Nguyen, and Nhat Ho. 2024. Neural collapse for cross-entropy class-imbalanced learning with unconstrained relu features model. In *Proceedings of the 41st International Conference on Machine Learning*, pages 10017–10040. PMLR.

Hien Dang, Tho Tran, Stanley Osher, Hung Tran-The, Nhat Ho, and Tan Nguyen. 2023. Neural collapse in deep linear networks: From balanced to imbalanced data. In *Proceedings of the 40th International Conference on Machine Learning*, pages 6873–6947. PMLR.

Zhongfen Deng, Hao Peng, Dongxiao He, Jianxin Li, and Philip S. Yu. 2021. HTCInfoMax: A global model for hierarchical text classification via information maximization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3259–3265, Online. 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: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jintong Gao, He Zhao, Dandan Guo, and Hongyuan Zha. 2024. Distribution alignment optimization through neural collapse for long-tailed classification. In *Proceedings of the 41st International Conference on Machine Learning*, pages 14969–14987. PMLR.

Gene H. Golub and Charles F. Van Loan. 2013. *Matrix Computations*, 4th edition. The Johns Hopkins University Press, Baltimore, MD.

Siddharth Gopal and Yiming Yang. 2013. Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 257–265, Chicago, Illinois, USA. Association for Computing Machinery.

Ke Ji, Yixin Lian, Jingsheng Gao, and Baoyuan Wang. 2023. Hierarchical verbalizer for few-shot hierarchical text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 2918–2933, Toronto, Canada. Association for Computational Linguistics.

Gibaeg Kim, SangHun Im, and Heung-Seon Oh. 2024. Hierarchy-aware biased bound margin loss function for hierarchical text classification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7672–7682, Bangkok, Thailand. Association for Computational Linguistics.

Fanshuang Kong, Richong Zhang, and Ziqiao Wang. 2025. LH-Mix: Local hierarchy correlation guided mixup over hierarchical prompt tuning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 636–646, Toronto ON, Canada. Association for Computing Machinery.

Kamran Kowsari, Donald E. Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S. Gerber, and Laura E. Barnes. 2017. HDLTex: Hierarchical deep learning for text classification. In *Proceedings of the 16th IEEE International Conference on Machine Learning and Applications*, pages 364–371, Cancun, Mexico. IEEE.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Qitong Liu, Hao Peng, Xiang Huang, Zhifeng Hao, Qingyun Sun, Zhengtao Yu, and Philip S. Yu. 2025. Hierarchical text classification optimization via structural entropy and singular smoothing. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5283–5297.

Yuning Mao, Jingjing Tian, Jiawei Han, and Xiang Ren. 2019. Hierarchical text classification with reinforced label assignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 445–455, Hong Kong, China. Association for Computational Linguistics.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814, Haifa, Israel. Omnipress.

Vardan Papyan, X. Y. Han, and David L. Donoho. 2020. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Thanh Duc Pham, Nam Le Hai, Linh Ngo Van, Diep Thi-Ngoc Nguyen, Sang Dinh, and Thien Huu Nguyen. 2025. Mitigating non-representative prototypes and representation bias in few-shot continual relation extraction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 10791–10809, Vienna, Austria. Association for Computational Linguistics.

Evan Sandhaus. 2008. The new york times annotated corpus. Linguistic Data Consortium, Philadelphia, PA. LDC2008T19.

Aixin Sun, Ee-Peng Lim, Wee-Keong Ng, and Jaideep Srivastava. 2004. Blocking reduction strategies in hierarchical text classification. *IEEE Transactions on Knowledge and Data Engineering*, 16(10):1305–1308.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada. OpenReview.net.

Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. 2022a. Incorporating hierarchy into text encoder: A contrastive learning approach for hierarchical text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 7109–7119, Dublin, Ireland. Association for Computational Linguistics.

Zihan Wang, Peiyi Wang, Tianyu Liu, Binghuai Lin, Yunbo Cao, Zhifang Sui, and Houfeng Wang. 2022b. HPT: Hierarchy-aware prompt tuning for hierarchical text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3740–3751, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sishi Xiong, Yu Zhao, Jie Zhang, Mengxiang Li, Zhongjiang He, Xuelong Li, and Shuangyong Song. 2024. Dual prompt tuning based contrastive learning for hierarchical text classification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12146–12158, Bangkok, Thailand. Association for Computational Linguistics.

Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1):69–90.

Chao Yu, Yi Shen, and Yue Mao. 2022. Constrained sequence-to-tree generation for hierarchical text classification. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1865–1869, Madrid, Spain. Association for Computing Machinery.

Simon Chi Lok Yu, Jie He, Víctor Gutiérrez-Basulto, and Jeff Z. Pan. 2023. Instances and labels: Hierarchy-aware joint supervised contrastive learning for hierarchical multi-label text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8858–8875, Singapore. Association for Computational Linguistics.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, Online. Association for Computational Linguistics.

Juncheng Zhou, Lijuan Zhang, Yachen He, Rongli Fan, Lei Zhang, and Jian Wan. 2025. A novel negative sample generation method for contrastive learning in hierarchical text classification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5645–5655, Abu Dhabi, UAE. Association for Computational Linguistics.

He Zhu, Junran Wu, Ruomei Liu, Yue Hou, Ze Yuan, Shangzhe Li, Yicheng Pan, and Ke Xu. 2024. HILL: Hierarchy-aware information lossless contrastive

learning for hierarchical text classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4731–4745, Mexico City, Mexico. Association for Computational Linguistics.

A Dataset statistics

Dataset	$\lvert\mathcal{Y}\rvert$	D	Avg($\lvert y_i\rvert$)	Train	Val	Test
WOS	141	2	2.00	30070	7518	9397
RCV1-V2	103	4	3.24	20833	2316	781265
NYT	166	8	7.60	23345	5834	7292

Table 6: Statistics of the three HTC datasets. $\lvert\mathcal{Y}\rvert$ denotes the total number of labels. D denotes the depth of label hierarchy. $\mathrm{Avg}(\lvert y_i\rvert)$ denotes the mean count of labels assigned to each sample.

B Implementation Details

B.1 Prompt Tuning-based Model for HTC

The prompt setting in HTC is generally categorized into hard and soft prompts. Hard prompts employ discrete text templates designed manually. For instance, for a label hierarchy of depth 2, the input can be formulated as [CLS] X [SEP] It belongs to [Mask1] [Mask2] [SEP] (Xiong et al., 2024), where [Maski] is used to predict the label at the i-th level. In contrast, soft prompts utilize continuous, learnable vectors as virtual template words, formatting the input as [CLS] X [SEP] [V1] [Mask1] [V2] [Mask2] [SEP] (Wang et al., 2022b; Liu et al., 2025). Here, virtual tokens [Vi] are trainable parameters designed to encapsulate hierarchical information for the i-th level and are optimized jointly with the model.

In this work, we follow the architecture proposed by Wang et al. (2022b), which reframes classification as a masked language modeling task using soft prompts. Critically, we adhere to its weight-tying scheme, where our label embedding matrix—the geometric prototype constructed by HGS—plays a dual role: (1) it is used to initialize the input soft prompt tokens and (2) its transpose serves as the weights of the final masked language modeling head, which calculates the classification logits at the [Maski] position.

B.2 Hyperparameter Settings

We have listed the hyperparameter settings for the three datasets in Table 7.

Hyperparameter	θ_t	γ	λ
WOS	20	0.90	0.01
RCV1-V2	10	0.90	1.0
NYT	25	0.95	0.1

Table 7: Hyperparameter settings.

C Experimental Results

C.1 Research on Feature Space of Labels

To evaluate the effectiveness of LGSA in optimizing the geometric structure of labels in the feature space, a series of geometric metrics assessments are conducted on the label embeddings learned by LGSA and baseline models (HPT and SIHTC). As shown in Table 8, the experimental results reveal significant differences between the models in constructing the geometric structure of feature space, which stem from their inherent mechanisms.

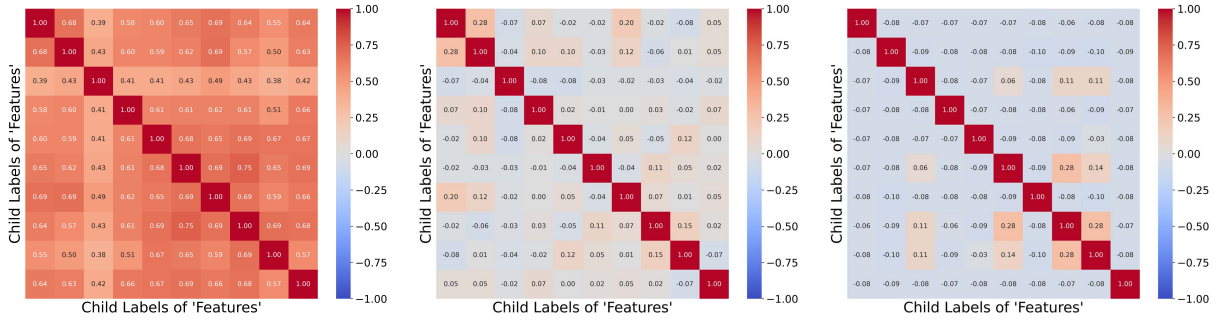
Orthogonality Analysis. LGSA demonstrates a superior ability to construct a separable feature space. Experimental results across three datasets show that the head label and sibling residual orthogonality metrics of LGSA consistently approach zero, significantly outperforming HPT and SIHTC. Visual evidence in Figure 5(a) further confirms this: while HPT (Figure 5(a)) shows strong off-diagonal correlations and SIHTC (Figure 5(b)) exhibits residual dependencies, the LGSA heatmap (Figure 5(c)) closely approximates an identity matrix. These results indicate that through HGS initialization and HGA regularization, LGSA effectively ensures global separation between top-level branches and preserves the independence of sibling-specific information.

Parent-Child Consistency. LGSA achieves high parent-child label similarity, aligning with its objective to model "is-a" hierarchies by explicitly constraining the direction of parent and child vectors via a dedicated similarity loss. In contrast, SIHTC yields significantly lower similarity on this metric. This divergence stems from their underlying strategies: while SIHTC employs singular spectrum smoothing—a global regularization aimed at mitigating representation degeneration—it lacks local geometric constraints for specific hierarchical pairs. These results demonstrate that LGSA successfully injects local parent-child priors, whereas SIHTC focuses primarily on the global spectral properties of the embedding space.

Within-Class Variance. Analysis of average within-class variance reveals a distinct characteristic of LGSA. LGSA exhibits slightly higher variance than the baselines, a direct result of its structural constraints. While baselines tend to collapse all samples sharing a head label into a single region, LGSA employs sibling residual orthogonality

Dataset	Model	Head Orth.	P-C Sim.	Norm Ratio.	Intra Var.	Sib. Orth.
WOS	HPT (Wang et al., 2022b)	0.7966	0.7029	0.9988	67.3552	0.5727
	SIHTC (Liu et al., 2025)	0.0852	0.0749	1.0173	70.9116	0.0497
	LGSA (Proposed)	0.0743	0.7837	0.9988	79.6630	0.0314
NYT	HPT (Wang et al., 2022b)	0.6328	0.7359	0.9997	132.2867	0.5463
	SIHTC (Liu et al., 2025)	0.1996	0.3163	1.0008	150.3072	0.0540
	LGSA (Proposed)	0.0037	0.8339	1.0059	173.1808	0.0505
RCV1-V2	HPT (Wang et al., 2022b)	0.6620	0.7123	0.9999	104.2893	0.6065
	SIHTC (Liu et al., 2025)	0.1684	0.2130	0.9995	126.5298	0.0497
	LGSA (Proposed)	0.0079	0.9063	1.0019	124.6783	0.0435

Table 8: Quantitative analysis of geometric structure for the learned feature space. Note: **Head Orth.**: Head label orthogonality; **P-C Sim.**: Parent-child label similarity; **Norm Ratio.**: Norm decay ratio; **Intra Var.**: Average within-class variance; **Sib. Orth.**: Orthogonality of sibling label residuals.



(a) HPT (b) SIHTC (c) LGSA

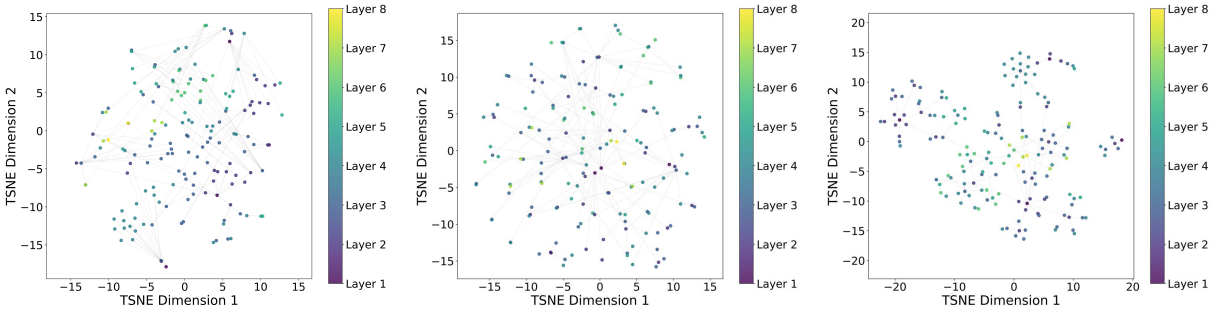
Figure 5: Residual similarity (cosine similarity) heatmaps of sibling label residual embeddings for the child labels under the parent label 'Features'.

to map child labels to distinct sub-regions within the parent space. This preservation of fine-grained structural information naturally increases the overall variance of the parent label. This is corroborated by the t-SNE visualization in Figure 6. In contrast to the unstructured agglomerations of HPT (Figure 6(a)) and SIHTC (Figure 6(b)), LGSA (Figure 6(c)) forms independent clusters with clear local tree-like geometries. These clusters correspond to distinguishable child label sub-regions, visually confirming the effective modeling of hierarchical granularity.

Norm Decay Analysis. We observe a universal phenomenon across LGSA and all baselines: the norm decay ratio consistently approximates 1.0. This suggests that the norm-sample count dependency—as predicted by imbalanced neural collapse theory—does not significantly manifest in these prompt-based fine-tuning architectures. We attribute this to two factors. First, the classification loss implicitly incentivizes the model to maintain label embedding norms to ensure the discriminability of long-tailed labels, counteracting potential decay. Second, these models prioritize optimizing the angular geometry to achieve discriminability. Once an effective angular structure is established, the influence of embedding norms on the classification decision diminishes. Consequently, the models converge to a solution that satisfies angular constraints while maintaining uniform norms. This reveals that in hierarchical classification, angular structuring is the primary mechanism for feature space organization rather than norm modulation.

C.2 Research on Low-resource Settings

To further validate the effectiveness of LGSA, the experiment in a low resource setting is conducted. Consistent with previous research (Wang et al., 2022a), 10% of the data is randomly selected from the training set for training. The experimental results, shown in Table 9, show that LGSA outperforms the baseline models HPT and SIHTC on all three datasets, except for the Macro-F1 metric on the NYT dataset. These results demonstrate that when supervisory signals are insufficient, this structured prior knowledge can serve as an effective regularizer, guiding and constraining the learning direction of the feature space and compensating



(a) HPT (b) SIHTC (c) LGSA

Figure 6: TSNE visualization of label embedding spaces learned by (a) HPT, (b) SIHTC, and (c) LGSA, where colors denote label hierarchy levels.

Model	WOS		RCV1-V2		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
HPT (Wang et al., 2022b)	<u>82.34</u>	<u>73.29</u>	<u>84.34</u>	57.14	75.42	58.18
SIHTC (Liu et al., 2025)	81.72	73.09	84.20	<u>57.53</u>	<u>75.51</u>	**59.51**
LGSA (proposed)	**82.36**	**73.71**	**84.37**	**58.00**	**75.60**	<u>59.28</u>

Table 9: Experimental results on three HTC datasets. In each column, the best value is in bold, the second best value is underlined.

Models	RCV1-V2		NYT	
	C-MiF1	C-MaF1	C-MiF1	C-MaF1
HPT (Wang et al., 2022b)	87.00	67.91	79.28	68.26
SIHTC (Liu et al., 2025)	87.48	69.88	79.77	68.89
LGSA (proposed)	**87.53**	**70.25**	**79.96**	**69.43**

Table 10: Evaluation results of hierarchical consistency.

for the information loss caused by data scarcity. This enables LGSA to learn more generalizable label representations even in low-resource settings, thereby maintaining its performance advantage.

C.3 Research on Hierarchical Consistency

The hierarchical consistency performance of a model directly reflects its capability to learn the hierarchical label structure. To this end, the hierarchical consistency metrics C-MicroF1 (C-MiF1) and C-MacroF1 (C-MaF1) (a label is considered correctly classified only if the leaf label and all of its ancestor labels are accurately predicted together; Yu et al. (2022)) are employed to evaluate the predictions of LGSA and the baselines (HPT and SIHTC). As shown in Table 10, LGSA achieved the best performance on all hierarchical consistency metrics for RCV1-V2 and NYT datasets. This result strongly demonstrates that the predicted labels generated by LGSA can better follow the hierarchical constraint that if a child node is predicted, its parent node must also be predicted. This superior hierarchical consistency performance is mainly attributed to the fact that LGSA, through the explicit parent-child similarity loss and sibling residual orthogonality constraint, forces the model to learn and reconstruct the "is-a" inheritance relationship and local topological structure of labels in the feature space.