Ensembling Prompting Strategies for Zero-Shot Hierarchical Text Classification with Large Language Models

Mingxuan Xia[12*], Zhijie Jiang[1*], Haobo Wang[12†], Junbo Zhao[23], Tianlei Hu[3], Gang Chen[3]
[1]School of Software Technology, Zhejiang University
[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[3]College of Computer Science and Technology, Zhejiang University
{xiamingxuan,zjjjj882,wanghaobo,j.zhao,htl,cg}@zju.edu.cn

Abstract

Hierarchical text classification aims to classify documents into multiple labels within a hierarchical taxonomy, making it an essential yet challenging task in natural language processing. Recently, using Large Language Models (LLM) to tackle hierarchical text classification in a zero-shot manner has attracted increasing attention due to their cost-efficiency and flexibility. Given the challenges of understanding the hierarchy, various HTC prompting strategies have been explored to elicit the best performance from LLMs. However, our empirical study reveals that LLMs are highly sensitive to these prompting strategies—(i) within a task, different strategies yield substantially different results, and (ii) across various tasks, the relative effectiveness of a given strategy varies significantly. To address this, we propose a novel ensemble method, HiEPS, which integrates the results of diverse prompting strategies to promote LLMs' reliability. We also introduce a path-valid voting mechanism for ensembling, which selects a valid result with the highest path frequency score. Extensive experiments on three benchmark datasets show that HiEPS boosts the performance of single prompting strategies and achieves SOTA results. The source code is available at https://github.com/MingxuanXia/HiEPS.

1 Introduction

Hierarchical Text Classification (HTC) (Sun and Lim, 2001) is a significant but challenging task in Natural Language Processing (NLP) that aims to assign multiple labels to a document within a hierarchical taxonomy. Unlike standard text classification with a flat and limited label space, HTC deals with complex label hierarchies, where higher-level labels represent broader concepts, while lower-level ones capture more specific subtopics. In recent years, HTC has attracted growing attention from both academia and industry (Zhang et al., 2025; Tabatabaei et al., 2024), driven by its practical relevance in real-world applications such as document organization (Kowsari et al., 2017), product categorization in e-commerce (McAuley and Leskovec, 2013), and information retrieval (Lehmann et al., 2015).

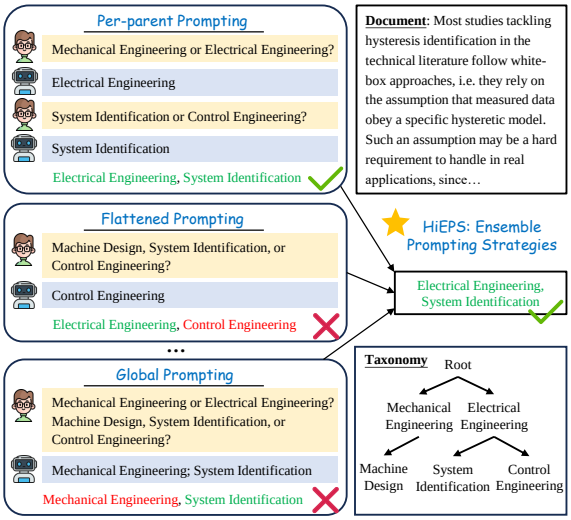


Figure 1: An example of zero-shot HTC using GPT-3.5 on WOS, where the ensemble of diverse prompting strategies offers a stable and robust solution.

Traditional HTC methods (Zhou et al., 2020; Chen et al., 2021; Wang et al., 2022b) typically train models on large amounts of labeled data within a static taxonomy, rendering them resource-intensive for data collection and re-training when the taxonomy changes over time. Therefore, recent studies have increasingly focused on zero-shot HTC (Bhambhoria et al., 2023; Bongiovanni et al., 2023; Paletto et al., 2024), where classification is performed through pre-trained language models (Li et al., 2024) without access to any labeled training data. In particular, large language models (LLMs)

[*] Equal contribution.
[†] Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 18189–18208
November 4-9, 2025

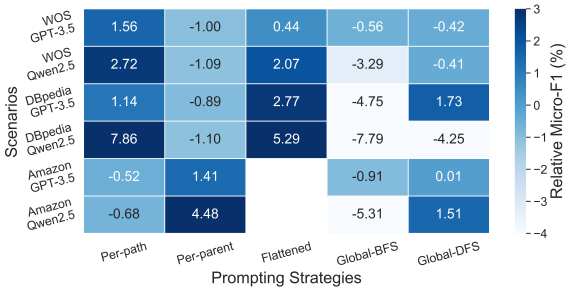


Figure 2: The relative Micro-F1 of different prompting strategies (see section 2.2 for details) under different scenarios, showing LLMs are highly sensitive to them. Given a task and an LLM, the relative Micro-F1 of a prompting strategy is calculated by subtracting the average Micro-F1 of all strategies from its own.

(Zhao et al., 2023) exhibit considerable promise for zero-shot HTC, owing to their advanced capabilities in language comprehension and generation.

In this work, we focus on *zero-shot HTC using LLMs*, where the complexity of capturing the hierarchical taxonomy makes prompt design one of the most crucial issues in determining performance. So far, different prompting strategies have been explored to tackle the challenges of HTC, including flattened (Bhambhoria et al., 2023), per-parent (Chen et al., 2024), or global (Zhang et al., 2025) prompting. While most existing works study these strategies individually, our empirical studies reveal that, **LLMs are highly sensitive to prompting strategy for HTC**——as shown in Figure 2, (i) within a single scenario, different prompting strategies yield substantially different results; and (ii) across different scenarios, the relative effectiveness of a given strategy varies significantly. This indicates that relying on the result from one single prompting strategy often leads to unstable and unreliable performance in zero-shot HTC.

To address the instability of single outputs from LLMs, one of the most widely adopted strategies is ensembling. In particular, self-consistency (Wang et al., 2023) integrates the results from different decoding paths and demonstrates superiority on hard tasks, such as reasoning and code generation (Chen et al., 2023). Nevertheless, this sampling-based ensemble method is still confined to a single HTC prompting strategy, thus remaining vulnerable to LLMs' prompt sensitivity, see discussion in section 3.3. To this end, we propose a novel ensemble framework for HTC called **HiEPS**, which ensembles the results of diverse prompting strategies. As shown in the example in Figure 1, by combining results based on diverse interpretations of the label hierarchy, HiEPS realizes the mutual complementarity of the advantages offered by these strategies, thus improving LLMs' stability and robustness for HTC. Specifically, given the results from different prompting strategies, we introduce a *path-valid voting* mechanism for integration that selects a valid label path with the highest path frequency score, which not only ensures the trustworthiness but also improves the performance against majority voting.

Empirically, we validate the effect of HiEPS on three benchmark datasets and various LLMs, showing that HiEPS largely improves single prompting strategies and establishes state-of-the-art performance on HTC tasks. Besides, our cost analysis indicates that although using multiple prompting strategies makes HiEPS more computationally expensive, it can achieve significant improvements, and we also introduce some options to reduce resource consumption; see Appendix A for details.

2 The Proposed Method

2.1 Preliminaries

Given a taxonomy structure $\mathcal{T} = (\mathcal{Y}, \mathcal{E}, \mathcal{R})$, which commonly takes the form of a tree, Hierarchical Text Classification (HTC) aims to classify a document x into its label path $\boldsymbol{y}$. Specifically, $\mathcal{Y} = \bigcup_{l=1}^{L} \mathcal{Y}_l$ denotes the node (label) set, where L is the number of layers in the taxonomy structure and $\mathcal{Y}_l$ denotes the label set of layer l. $\mathcal{E}$ represents the set of edges directed from the parent nodes to their children nodes, and $\mathcal{R}$ denotes the root node, i.e., the parent node of $\mathcal{Y}_1$. The taxonomy path $\boldsymbol{y} = [y_1, \ldots, y_L]$ is a list of L labels, where $y_l \in \mathcal{Y}_l$ denotes its corresponding label for layer l. Besides, the taxonomy structure can also be a directed acyclic graph (DAG), where one node can have multiple parents with different meanings, which is more realistic. In this study, we focus on **zero-shot HTC using large language models (LLMs)**, where we verbalize the label taxonomy and prompt LLMs to classify the input documents.

2.2 Diverse Prompting Strategies

Due to the inherent complexity of hierarchical structures, taxonomies can be verbalized in various ways through different prompting strategies. In this paper, we explore a range of prompting strategies that encompass diverse perspectives in interpreting the taxonomy, including *local strategies* that transform HTC into simpler sub-tasks, and *global*

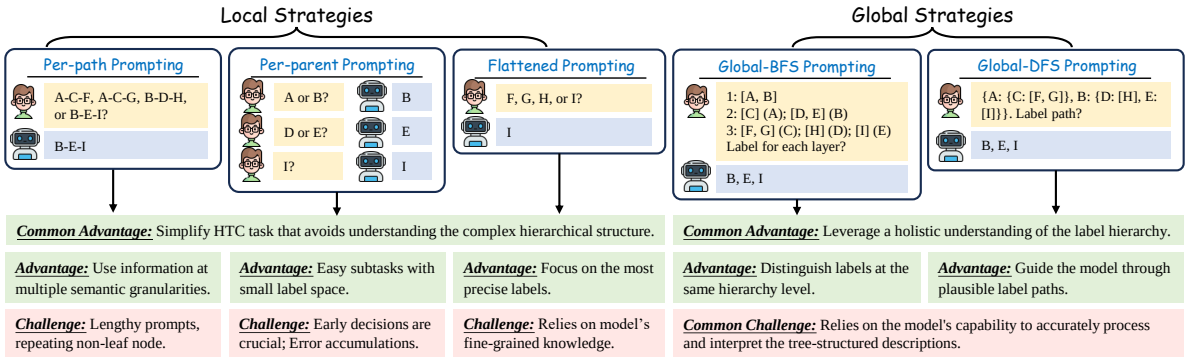


Figure 3: Examples, advantages, and challenges of different prompting strategies. The taxonomy structure follows the hierarchical tree in Figure 4.

strategies that emphasize a holistic understanding of global information. A detailed overview of these strategies, including their respective advantages and limitations, is presented as follows:

Per-path prompting is a local strategy that simplifies HTC into label path classification, aiming to directly identify the correct label path from all possible paths in the taxonomy. This method simplifies the model's understanding of complex hierarchies and mitigates the uncertainty in decision-making across different layers. Nonetheless, the repeated inclusion of non-leaf node descriptions often leads to lengthy prompts, which can adversely impact the model's ability to process and understand the input effectively.

Per-parent prompting is a local strategy that reformulates HTC as a series of parent-node classification tasks. Specifically, starting from the root node as the initial parent, the model classifies the given sample into one of the current parent's child nodes. The selected child node then becomes the new parent, and this process continues recursively until a leaf node is reached. This method aligns with the generation process of the label hierarchy and benefits from a reduced decision space at each step, which can notably improve classification accuracy. However, this method is prone to error accumulation, as incorrect decisions at earlier stages can propagate through subsequent classifications.

Flattened prompting is a local strategy that aims to directly identify the correct label from all leaf nodes in the taxonomy. The corresponding taxonomy path is then recovered by propagating the selected leaf node to its ancestors. Note that we only apply this strategy for tasks with tree structures, since tasks with DAG structures may incur ambiguity during propagation. By concentrating

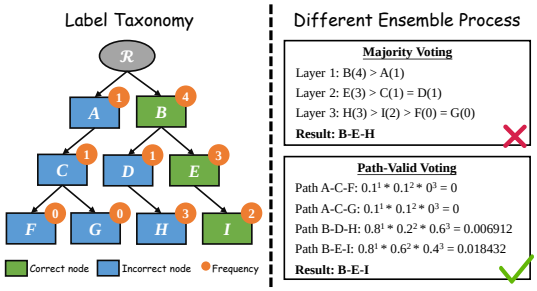


Figure 4: An example of using majority voting and path-valid voting for ensembling.

on the most fine-grained labels, this approach helps the model bypass a complex understanding of intermediate hierarchical levels. Nevertheless, it heavily challenges the model's capability to accurately understand and differentiate nuanced semantic variations among fine-grained labels.

Two global strategies, **Global-BFS** and **Global-DFS**, first present the full taxonomy structure using breadth-first search (BFS) and depth-first search (DFS), respectively, and then query the LLM to select the correct labels for the given sample. These approaches leverage the full label hierarchy, allowing the model to thoroughly understand inter-label relationships and hierarchical dependencies. Specifically, BFS facilitates better differentiation of semantic meanings among labels at the same hierarchy level, while DFS excels at guiding the model through plausible label paths. The core challenge of this strategy lies in its reliance on the model's capability to accurately process and interpret the tree-structured descriptions.

2.3 HiEPS: Ensembling Prompting Strategies

Our preliminary study illustrated in Figure 2 shows that LLMs are highly sensitive to prompting strate-

Table 1: Ratios of invalid paths using majority voting.

GPT-3.5			Qwen2.5		
WOS	DBpedia	Amazon	WOS	DBpedia	Amazon
13.5%	12.4%	14.9%	10.5%	7.6%	24.2%

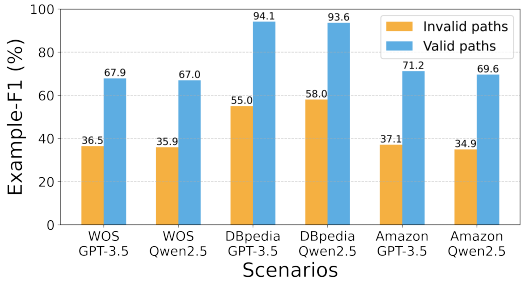


Figure 5: Comparison of Example-F1 scores between majority voting results with valid paths and invalid ones.

gies in HTC, implying that relying on a single strategy often leads to unreliable results. This is because the strengths and limitations of each strategy manifest differently when applied across various tasks or LLMs. To address this, we propose a simple yet effective method **HiEPS**, which ensembles the results of diverse prompting strategies to combine their respective strengths.

Formally, after prompting the LLM with different strategies, we obtain a list of predicted label paths $\hat{\mathcal{P}}$. To ensemble these label paths into a final prediction $\hat{\boldsymbol{y}}$, we propose to conduct layer-wise voting and reformulate $\hat{\mathcal{P}}$ into $\{S_l\}_{l=1}^{L}$, where S_l represents the list of predicted labels for layer l. An intuitive way is to conduct majority voting individually on each layer, namely:

$$\hat{y}_l = \arg\max_{y \in \mathcal{Y}_l} \#\{i | S_l^{(i)} = y\} \tag{1}$$

where $\#\{\cdot\}$ counts the number of elements in the given set and $S_l^{(i)}$ denotes the i-th element in S_l.

However, voting for each layer independently may lead to results with invalid label paths, see the statistics in Table 1 and the example in Figure 4. This not only compromises the method's trustworthiness but also undermines the method's effectiveness, where Figure 5 demonstrates that the majority voting results with invalid paths suffer from substantial degradation. The primary cause leading to such invalid paths is that, given the complex and large-scale label structure in HTC, LLMs sometimes fail to produce a valid path under a single prompting strategy (see discussion in section 3.3), and consequently reach invalid ensembled results.

To this end, we propose a novel *path-valid voting* method for ensembling, which selects a valid label path with the highest path frequency. Specifically, the path frequency is defined as the product of the frequencies of each label in the path:

$$\text{PathFrequency}(\boldsymbol{y}) = \prod_{l=1}^{L} \left(\frac{\#\{i | S_l^{(i)} = y_l\}}{|S_l|} \right)^l \tag{2}$$

where $|S_l|$ denotes the length of label list S_l. Note that the frequencies are powered by their layer numbers to make the result more focused on the fine-grained labels. The final results of path-valid voting are then formalized as:

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y} \in \mathcal{P}} \text{PathFrequency}(\boldsymbol{y}) \tag{3}$$

where $\mathcal{P}$ denotes the set of all valid paths in the hierarchical taxonomy.

Remark. Although HiEPS and Self-Consistency (SC) (Wang et al., 2023) both first generate multiple results and then integrate them, they have notable distinctions: (i) HiEPS ensembles multiple prompting strategies, while SC relies on single prompting strategies and performs ensembling by repeatedly sampling. In essence, given the fact that LLMs are highly sensitive to the choice of HTC prompting strategies, the superiority of HiEPS stems from the ability to incorporate **multi-perspective structural understanding** for ensemble—something SC is inherently incapable of achieving on its own; (ii) HiEPS adopts a path-valid voting mechanism during the ensemble, which turns out to be more trustworthy and effective than simple majority voting as implemented in SC. See section 3 and Appendix B.1 for more experimental analysis.

2.4 Multi-answer Prompting

In addition to the prompting strategies introduced in section 2.2 that have the LLM output a single answer, we also investigate *multi-answer prompting* strategies, which allow the LLM to propose multiple possible answers, improving the recall of the correct labels[1]. Specifically:

Per-path-multi prompting queries LLMs to output all possible label paths from the taxonomy.

Per-parent-multi prompting follows per-parent prompting but allows the LLM to output multiple possible labels for each step. During the process,

[1]The results of how multi-answer prompting improves the recall of HiEPS can be found in Table 4.

Table 2: Comparisons of Micro-F1 (%), Macro-F1 (%), and Example-F1 (%) using **GPT-3.5**. Average results over three runs are reported. The best result is bold and the second best is underlined. ↑ means the improvement of HiEPS over the **best strategy** (marked in red) of the eight ensembled strategies introduced in section 2.

Method	WOS			DBpedia			Amazon		
	Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1
BART-NLI	57.42	52.32	57.42	71.60	62.76	71.60	22.83	8.45	22.83
BART-NLI+LLM	61.95	51.83	61.92	85.33	79.38	85.20	35.96	15.96	34.29
UP	62.90	49.41	62.90	78.14	63.38	78.14	38.68	13.86	38.69
HiLA+UP	57.74	43.92	57.74	79.03	65.37	79.03	44.16	16.79	44.16
Per-path	59.77	50.53	58.62	80.43	70.00	79.70	60.37	24.73	59.96
Per-parent	57.21	45.52	56.14	78.40	64.10	76.41	62.30	25.86	61.91
Flattened	58.65	47.23	55.02	82.06	75.67	78.93	-	-	-
Global-BFS	57.65	48.09	56.63	74.54	66.34	73.48	59.98	22.50	59.49
Global-DFS	57.79	49.02	56.79	81.02	71.64	79.57	60.90	24.11	60.39
Per-path-multi	46.66	42.02	51.33	61.68	55.99	67.43	50.64	19.16	56.48
Per-parent-multi	48.96	47.66	50.35	63.43	61.27	65.65	55.07	23.97	57.60
Flattened-multi	50.68	47.47	51.81	69.02	66.08	71.58	-	-	-
CoT	57.73	45.50	57.08	82.20	73.68	80.51	64.24	26.86	63.66
SC with mv	58.88	45.49	58.41	84.62	76.40	83.43	65.64	26.87	64.92
SC with pvv	58.80	44.38	58.44	88.72	77.44	86.43	65.71	26.98	65.51
ToT	60.37	50.13	58.73	81.50	69.57	80.06	61.84	26.60	58.65
HiEPS	**65.41** ↑5.6	**55.11** ↑4.6	**65.39** ↑6.8	**91.87** ↑9.8	**84.61** ↑8.9	**91.87** ↑12.2	**67.07** ↑4.8	**29.99** ↑4.1	**67.01** ↑5.1

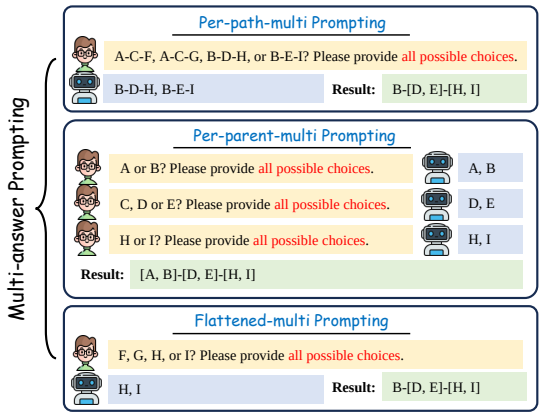


Figure 6: Examples of multi-answer prompting. The taxonomy follows the hierarchical tree in Figure 4.

the chosen nodes become the new parent nodes, and the union of their children is then used as the candidate label set for the next step.

Flattened-multi prompting queries LLMs to output all possible labels from all the leaf nodes in the taxonomy tree. Like flattened prompting, the results are obtained by propagating the selected leaf nodes to their ancestors, and it is only applied for tree-structured tasks.

Overall, the aforementioned local, global, and multi-answer prompting strategies are used for the ensemble in HiEPS. Figure 3 and 6 show examples for each of them. Note that though the results of multi-answer prompting are not label paths but rather a subtree of the taxonomy, we directly use this subtree for evaluation as well as frequency calculation when ensembling. Besides, HiEPS is not limited to these strategies and can be further extended by incorporating others.

3 Experiments

3.1 Setup

Datasets and Evaluation Metrics. We conduct experiments on the following three public datasets: 1) Web Of Science[2] (**WOS**) (Kowsari et al., 2017) consists abstracts of research papers which are labeled into a two-layer taxonomy tree, with 7 and 134 classes; 2) **DBpedia**[3] (Lehmann et al., 2015) consists of Wikipedia articles with a three-layer taxonomy tree, with 9/70/219 classes; 3) **Amazon**[4] (McAuley and Leskovec, 2013) consists of Amazon product reviews which are labeled into a three-layer taxonomy DAG, with 6/64/472 classes. For WOS and Amazon, we randomly sample 10% of the full datasets for evaluation, namely, 4698 and 5000 samples. For DBpedia, which contains hundreds of thousands of articles, we randomly sampled 5000 for evaluation. Following previous works (Chen et al., 2024; Zhang et al., 2025), we use Micro-F1, Macro-F1, and Example-F1 as the metrics for overall performance evaluation.

[2] https://huggingface.co/datasets/web_of_science

[3] https://www.kaggle.com/datasets/danofer/DBpedia-classes

[4] https://www.kaggle.com/datasets/kashnitsky/hierarchical-text-classification

Table 3: Comparisons of Micro-F1 (%), Macro-F1 (%), and Example-F1 (%) using **Qwen2.5-14B**.

Method	WOS			DBpedia			Amazon		
	Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1
BART-NLI	57.42	52.32	57.42	71.60	62.76	71.60	22.83	8.45	22.83
BART-NLI+LLM	56.64	44.96	56.63	73.90	68.23	73.83	33.17	12.00	32.44
UP	62.90	49.41	62.90	78.14	63.38	78.14	38.68	13.86	38.69
HiLA+UP	57.74	43.92	57.74	79.03	65.37	79.03	44.16	16.79	44.16
Per-path	62.27	49.95	61.50	86.83	77.19	86.60	53.24	19.68	52.67
Per-parent	58.46	48.07	58.38	77.87	63.14	76.22	58.40	24.40	57.21
Flattened	61.62	48.79	61.53	84.26	76.98	82.88	-	-	-
Global-BFS	56.26	42.84	55.57	71.18	60.18	68.31	48.61	14.31	47.25
Global-DFS	59.14	46.61	58.20	74.72	64.27	72.61	55.43	18.38	53.63
Per-path-multi	51.98	48.45	53.34	68.20	61.44	70.92	46.52	19.64	48.85
Per-parent-multi	46.68	46.73	47.51	60.85	51.82	61.60	47.90	19.83	46.87
Flattened-multi	51.42	46.12	51.98	65.55	55.14	67.82	-	-	-
CoT	57.34	42.94	57.17	83.69	70.89	82.53	60.70	20.86	58.25
SC with mv	58.57	44.83	58.47	86.42	74.98	85.67	62.44	21.71	59.87
SC with pvv	58.81	44.53	58.70	89.81	80.39	89.34	62.52	21.87	62.09
ToT	63.89	53.53	63.02	84.26	75.77	82.27	51.19	19.94	44.64
HiEPS	**65.01** ↑2.7	**53.80** ↑3.9	**65.01** ↑3.5	**92.33** ↑5.5	**85.07** ↑7.9	**92.33** ↑5.7	**62.59** ↑4.2	**25.38** ↑1.0	**62.46** ↑5.3

Baselines. For a comprehensive comparison, we exploit the following three types of baselines: 1) Zero-shot HTC methods. **BART-NLI** (Yin et al., 2019) transforms HTC into a textual entailment task and **BART-NLI+LLM** (Bhambhoria et al., 2023) first retrieves a few candidate labels through pre-trained entailment predictors and then uses LLMs for selection. **UP** (Bongiovanni et al., 2023) adopts pretrained embedding models to calculate the similarity between class names and leaf nodes, and then up-propagates their relevance scores to the full hierarchy for classification. **HiLA+UP** (Paletto et al., 2024) first leverages LLMs to create a deeper layer for the label hierarchy, and then adopts UP on the augmented taxonomy. 2) LLM for zero-shot HTC with **different prompting strategies**, as introduced in section 2.2 for details. 3) LLM for zero-shot HTC with advanced prompting techniques. Zero-shot **CoT** (Kojima et al., 2022), based on global-DFS prompting, employs chain-of-thought prompting by adding "Let's think step by step" before each answer. Self-Consistency (SC) (Wang et al., 2023) first samples diverse label paths using global-DFS[5] with CoT prompting and then generates the results using majority voting (**SC with mv**) or path-valid voting (**SC with pvv**). **ToT** (Yao et al., 2023), like per-parent-multi prompting, selects multiple possible children when classifying parent nodes, and then identifies the true label from all the selected leaf nodes where the label path leading to that node is treated as the final result.

[5]Global-DFS is adopted as the base strategy for SC (and CoT) since it offers the greatest potential for performance improvements. Please refer to Appendix B.1 for more details.

Implementation Details. In our main experiments, we exploit a closed-source model gpt-3.5-turbo-0125 (GPT-3.5) and an open-source model Qwen2.5-14B-Instruct (Qwen2.5) as the LLM for zero-shot HTC. We also investigate the effectiveness of HiEPS on more LLMs in section 3.4. We deploy open-source models on NVIDIA RTX A5000 GPUs, while the closed-source models are accessed through their official APIs. For the sampling-based method self-consistency, we sample the decoding path 10 times with a temperature of 0.8, and for other LLM generation processes, the temperature is set to a lower value of 0.5. For baseline BART-NLI and UP, we use BART-Large-MNLI (Yin et al., 2019) and mpnet-all (Reimers and Gurevych, 2019) as the pre-trained language model, following their original implementation.

3.2 Main Results

The comparison results of HiEPS with baselines using GPT-3.5 and Qwen2.5 are shown in Table 2 and Table 3. Overall, HiEPS outperforms all baselines on all tasks. For example, on DBpedia, HiEPS improves the Micro-F1, Macro-F1, and Example-F1 of the best baselines by margins of **3.15%**, **5.23%**, and **5.44%** when using GPT-3.5, and **2.52%**, **4.68%**, and **2.99%** when using Qwen2.5. The superior results in all scenarios imply the effectiveness of our proposed method.

Specifically, HiEPS largely improves the baselines based on pretrained entailment predictors (BART-NLI and BART-NLI+LLM) or embedding models (UP and HiLA+UP), especially on the hard

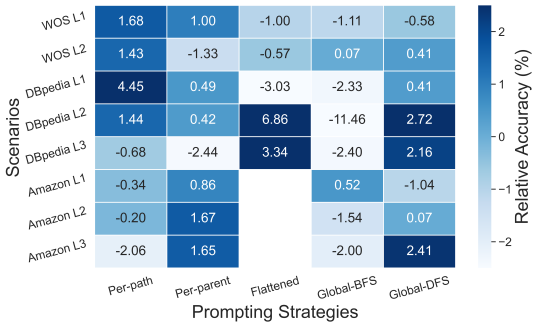


Figure 7: The relative layer-wise accuracy of different prompting strategies, where "L#" indicates the #-th layer. Given a layer of a task, the relative layer-wise accuracy of a prompting strategy is calculated by subtracting the average accuracy of all strategies from its own.

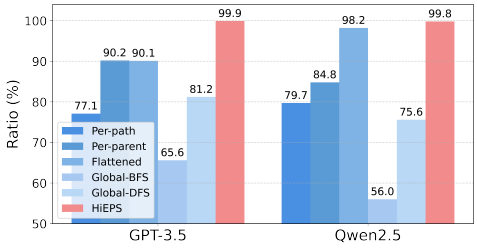


Figure 8: Comparison of the ratios of valid paths.

task, Amazon, where HiEPS with GPT-3.5 outperforms the best of them by margins of 22.91%, 13.20%, and 22.85% on the three metrics.

Also, HiEPS achieves significant improvements over individual prompting strategies. In particular, HiEPS improves the Example-F1 scores of the best prompting strategy on WOS, DBpedia, and Amazon by notable margins of **6.8%**, **12.2%**, and **5.1%** when using GPT-3.5, and **3.5%**, **5.7%**, and **5.3%** when using Qwen2.5. This indicates that HiEPS effectively mitigates the instability of individual prompting strategies, and the ensemble of them can lead to better performance.

In terms of advanced prompting techniques, CoT outperforms global-DFS prompting on DBpedia and Amazon by thinking step by step. By sampling and ensembling different decoding results, the self-consistency-based method achieves better performance than CoT, where SC with pvv outperforms SC with mv, indicating the significance of obtaining a valid label path. Moreover, ToT outperforms per-parent prompting on WOS and DBpedia by evaluating different child nodes. Even so, these advanced prompting methods still underperform HiEPS in all settings since they solely rely on one prompting strategy.

3.3 Further Analysis

In this subsection, we conduct further analysis to understand why the ensemble of diverse prompting strategies can boost performance.

Complementarity Effects of HiEPS. We show that different prompting strategies exhibit distinct strengths at varying levels of the hierarchy, making their integration achieve complementary effects to boost performance. Specifically, we compare the accuracies on each layer of the taxonomy when applying these strategies using GPT-3.5, and visualize the relative accuracy in Figure 7. As shown in the heat map, per-path prompting achieves the best result on the first layer of DBpedia, while flattened prompting achieves the best on the other two. Similarly, on Amazon, per-parent prompting achieves the best result on the first two layers while global-DFS achieves the best on the last one. These phenomena can be attributed to the different properties of individual strategy (as discussed in section 2.2). By making an ensemble of their strengths, HiEPS offers a robust HTC solution.

Comparison of the Ratios of Valid Paths. Another factor causing LLMs' instability for HTC is that using a single prompting strategy may result in invalid label paths (including generating labels outside the taxonomy, although such instances are relatively rare). As shown in Figure 8, most prompting strategies failed to produce valid paths over 10% of the time. This is because HTC tasks involve a large number of labels and complex semantic structures, making it difficult for LLMs to follow the hierarchical label relationships. Through the ensemble of multiple results with a path-valid voting mechanism, HiEPS effectively mitigates this issue, which hardly produces any invalid paths, thereby enhancing its stability and trustworthiness. It is worth noting that generating valid paths does not necessarily equal high performance. For example, even when applied with path-valid voting, SC with pvv still remains suboptimal in certain scenarios, such as WOS. This further highlights the importance of HiEPS in integrating diverse strategies.

Comparison to Self-Consistency. Moreover, we find that leveraging multiple prompting strategies leads to more diverse and informative results compared to SC, which uses multiple decoding results based on a single prompting strategy. In Table 4, we report the recalls and the average number of predicted labels (#La.) of self-consistency

Table 4: Comparisons of Recall and the average number of predicted labels (#La.) before ensemble.

Method	WOS GPT-3.5		WOS Qwen2.5		DBpedia GPT-3.5		DBpedia Qwen2.5		Amazon GPT-3.5		Amazon Qwen2.5	
	Recall	#La.	Recall	#La.	Recall	#La.	Recall	#La.	Recall	#La.	Recall	#La.
Self-consistency	68.02	20.0	69.84	20.0	92.25	30.0	94.26	30.0	74.05	30.0	71.25	30.0
HiEPS w/o multi	76.77	10.0	77.16	10.0	96.27	15.0	96.64	15.0	75.67	12.0	71.17	12.0
HiEPS	**91.46**	21.8	**87.67**	21.6	**98.41**	28.6	**98.74**	29.9	**81.00**	20.6	**78.68**	20.8

Table 5: Comparisons of Micro-F1 (%), Macro-F1 (%), and Example-F1 (%) using different LLMs on Amazon. The best result is bold and the second best is underlined. ↑ means the improvement of HiEPS over the best strategy.

Method	Llama3.1-8B-Instruct			Ministral-8B-Instruct			Qwen2.5-7B-Instruct			GPT-4o-mini		
	Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1
Per-path	49.16	14.77	48.91	38.81	11.12	37.55	49.44	15.88	48.85	59.42	23.52	58.67
Per-parent	53.78	20.96	53.47	48.47	15.16	45.28	53.01	18.77	51.34	64.17	27.12	62.83
Global-BFS	50.32	13.40	49.17	33.44	5.71	31.99	42.77	10.95	41.70	59.87	22.17	59.34
Global-DFS	50.96	13.46	50.00	35.82	9.05	32.94	51.58	14.99	50.76	61.18	24.38	60.69
Per-path-multi	42.22	13.78	44.50	36.04	8.57	36.35	37.66	9.97	39.54	53.23	21.89	53.98
Per-parent-multi	45.76	18.99	47.58	42.70	9.15	40.70	44.75	8.41	44.23	53.98	24.99	53.24
HiEPS	**59.25** ↑5.5	**22.60** ↑1.6	**59.24** ↑5.8	**53.23** ↑4.8	**17.59** ↑2.4	**51.98** ↑6.7	**56.40** ↑3.4	**20.64** ↑1.9	**55.99** ↑4.7	**65.85** ↑1.7	**29.04** ↑1.9	**65.75** ↑2.9

Method	Claude 3.5 Haiku			Gemini-2.0-Flash			GPT-4o			DeepSeek-R1		
	Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1
Per-path	54.70	19.25	53.53	69.82	33.76	69.62	67.67	32.84	66.18	67.83	29.95	67.80
Per-parent	64.43	28.27	62.87	66.72	30.73	66.14	69.04	31.70	67.53	68.69	32.80	68.67
Global-BFS	59.62	22.22	58.03	67.73	31.11	67.53	66.43	31.36	64.88	69.92	32.36	69.83
Global-DFS	58.06	20.27	55.77	68.91	31.81	68.59	69.34	33.16	68.03	70.12	32.44	70.04
Per-path-multi	57.83	27.05	61.15	66.07	31.92	67.94	61.23	27.29	64.49	64.42	30.09	67.50
Per-parent-multi	56.89	25.78	59.46	62.33	31.29	64.55	62.11	29.14	64.88	60.25	28.45	63.82
HiEPS	**66.49** ↑2.1	**30.24** ↑2.0	**66.38** ↑3.5	**70.71** ↑0.9	**34.01** ↑0.3	**70.70** ↑1.1	**70.69** ↑1.4	**34.05** ↑0.9	**70.09** ↑2.1	**71.57** ↑1.5	**34.10** ↑1.3	**71.57** ↑1.5

Table 6: Ablation studies of path-valid voting and multi-answer prompting in HiEPS.

Model	Method	WOS			DBpedia			Amazon		
		Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1	Micro-F1	Macro-F1	Example-F1
GPT-3.5	HiEPS with mv	63.65	**55.79**	63.63	89.27	82.88	89.27	66.15	**30.20**	66.12
	HiEPS w/o multi	63.19	52.85	63.07	90.79	83.84	90.72	65.76	28.52	65.69
	HiEPS	**65.41**	55.11	**65.39**	**91.87**	**84.61**	**91.87**	**67.07**	29.99	**67.01**
Qwen2.5	HiEPS with mv	63.71	**54.02**	63.71	90.88	83.87	90.88	61.26	24.76	61.23
	HiEPS w/o multi	63.72	52.59	63.72	90.93	82.85	90.93	62.10	24.97	61.44
	HiEPS	**65.01**	53.80	**65.01**	**92.33**	**85.07**	**92.33**	**62.59**	**25.38**	**62.46**

and HiEPS *before aggregating* the results, where HiEPS achieves much higher recalls on all tasks but a smaller number of labels on most tasks. We also report the results of *HiEPS w/o multi*, which do not adopt multi-answer prompting. Though producing much fewer labels, it still achieves higher recalls compared to self-consistency on most tasks.

3.4 Results on More LLMs

We further validate the effectiveness of HiEPS on more LLMs. Specifically, we compare HiEPS to each of the ensembled prompting strategies on Amazon using eight advanced LLMs: Three open-source LLMs, including Llama3.1-8B-Instruct, Ministral-8B-Instruct, and Qwen2.5-7B-Instruct; Four General-purpose closed-source LLMs, including GPT-4o-mini, Claude-3.5 Haiku, and Gemini-2.0-Flash, which are cost-effective, and GPT-4o, which is full-scaled; A reasoning-oriented LLM, DeepSeek-R1 (671B). As shown in Table 5, HiEPS demonstrates superiority across all the evaluated LLMs, which highlights its success in unifying the advantages of diverse prompting strategies.

3.5 Ablation Studies

In this section, we compare HiEPS with two of its variants: 1) *HiEPS with mv*, which uses majority voting instead of path-valid voting for ensembling, and 2) *HiEPS w/o multi*, which does not apply multi-answer prompting. As shown in Table 6, HiEPS improves the Micro-F1 and Example-F1 of HiEPS with mv on all scenarios. For Macro-F1, HiEPS achieves better results on half of the scenarios while suffering from decrements of 0.21%-0.68% on others. Despite this, HiEPS demonstrates an overall enhancement, where its improvements on the other two metrics surpass the minor reduction in Micro-F1. Moreover, SC with pvv also

achieves better performance than SC with mv on most tasks and metrics, see Table 2 and 3, which clearly shows the effectiveness of path-valid voting.

Compared to HiEPS w/o multi, HiEPS achieves superior results on all scenarios in all metrics. Specifically, when using GPT-3.5 on WOS, HiEPS outperforms HiEPS w/o multi by margins of 2.22%, 2.26%, and 2.32% in the three metrics. This stems from the significant improvements in recall (see Table 4 discussed above) when allowing LLMs to output multiple possible answers.

4 Related Work

4.1 LLM for Hierarchical Text Classification

Existing methods explore various ways to use LLMs for HTC. Bhambhoria et al. (2023) adopts LLMs to select the most relevant labels from the candidate labels retrieved by pre-trained entailment predictors. Chen et al. (2024) applies per-parent prompting with few-shot examples that are retrieved by the elaborately trained encoders. TELEClass (Zhang et al., 2025) proposed to query LLMs to produce core classes in the pruned taxonomy. Schmidt et al. (2024) proposed path-wise classification aiming to select the most relevant label path in the taxonomy. Despite directly adopting LLMs for prediction, HiLA (Paletto et al., 2024) leverages the LLM to create a deeper layer for the label hierarchy, achieving finer-grained similarity matching between documents and labels. In this paper, we reveal that LLMs are highly sensitive to prompting strategies and propose an ensemble framework to boost performance.

4.2 Ensemble Methods in LLM

Ensemble methods (Lakshminarayanan et al., 2017; Ganaie et al., 2022) aim to combine several individual models to obtain better generalization performance. In the era of LLMs, ensemble techniques have attracted increasing attention, where one of the most representative methods, self-consistency (Wang et al., 2023), aggregates the predictions sampled from multiple decoding paths, demonstrating notable effectiveness in many NLP tasks, such as math word problem (Cobbe et al., 2021; Shen et al., 2021b), reasoning (Wang et al., 2023; Weng et al., 2023), code generation (Chen et al., 2023), uncertainty estimation (Xiong et al., 2024), dialog system (Thoppilan et al., 2022), and hallucination mitigation (Zhang et al., 2024b). In addition to sampling-based ensembles, other methods aim to promote diversity by varying the inputs to LLMs, including using different prompt templates (Zhou et al., 2022; Zhang et al., 2024a), altering in-context examples (Lu et al., 2022; Pitis et al., 2023), or multi-lingual prompting (Qin et al., 2023). In this paper, we propose a novel ensemble method that combines diverse HTC prompting strategies, which is orthogonal to existing ensemble methods.

5 Conclusion

In this paper, we propose HiEPS, a novel ensemble method addressing the instability of LLMs in zero-shot HTC. HiEPS integrates diverse prompting strategies, enabling mutual compensation of their strengths, while the proposed path-valid voting guarantees the generation of valid label paths, improving both performance and reliability. Extensive experiments demonstrate that HiEPS outperforms individual strategies and achieves state-of-the-art results on benchmark datasets. We hope our work will inspire future research to explore ensemble techniques for tackling complex tasks.

Limitations

In this paper, we explore the ensemble of diverse prompting strategies for hierarchical text classification using large language models. While these strategies are intuitive and easy to implement, they can be further improved through techniques like label space pruning, taxonomy expansion, or incorporating more detailed descriptions of labels, and we leave the exploration of them to future work. Moreover, the proposed path-valid voting mechanism assumes that each level of the hierarchy has a single true label. When this assumption does not hold, e.g., in multi-label per-level settings with unknown label counts, this method requires further adaptation to remain effective.

Ethical Considerations

While the HTC datasets used in our paper are all publicly available and are widely adopted by researchers, utilizing LLMs for prediction may include bias and unfairness. Indeed, if HiEPS is used with such biased predictions, it may unpleasantly yield unfair and biased predictions based on characteristics like race, gender, disabilities, LGBTQ, or political orientation. To alleviate this issue, we recommend that potential users first use bias reduction and correction techniques to remove biased text

and predictions so as to improve overall fairness and ethical standards.

Acknowledgements

This paper is supported by the National Regional Innovation and Development Joint Fund (No. U24A20254). Haobo Wang is also supported by the Fundamental Research Funds for the Zhejiang Provincial Universities (No. 226-2025-00004).

References

Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical transfer learning for multi-label text classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6295–6300. Association for Computational Linguistics.

Rohan Bhambhoria, Lei Chen, and Xiaodan Zhu. 2023. A simple and effective framework for strict zero-shot hierarchical classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1782–1792. Association for Computational Linguistics.

Lorenzo Bongiovanni, Luca Bruno, Fabrizio Dominici, and Giuseppe Rizzo. 2023. Zero-shot taxonomy mapping for document classification. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, SAC 2023, Tallinn, Estonia, March 27-31, 2023*, pages 911–918. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4370–4379. Association for Computational Linguistics.

Huiyao Chen, Yu Zhao, Zulong Chen, Mengjia Wang, Liangyue Li, Meishan Zhang, and Min Zhang. 2024. Retrieval-style in-context learning for few-shot hierarchical text classification. *Trans. Assoc. Comput. Linguistics*, 12:1214–1231.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation. *CoRR*, abs/2311.17311.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

M. A. Ganaie, Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N. Suganthan. 2022. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.*, 115:105151.

Siddharth Gopal and Yiming Yang. 2013. Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*, pages 257–265. ACM.

Wei Huang, Enhong Chen, Qi Liu, Yuying Chen, Zai Huang, Yang Liu, Zhou Zhao, Dan Zhang, and Shijin Wang. 2019. Hierarchical multi-label text classification: An attention-based recurrent network approach. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1051–1060. ACM.

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 (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2918–2933. Association for Computational Linguistics.

Ke Ji, Peng Wang, Wenjun Ke, Guozheng Li, Jiajun Liu, Jingsheng Gao, and Ziyu Shang. 2024. Domain-hierarchy adaptation via chain of iterative reasoning for few-shot hierarchical text classification. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 6315–6323. ijcai.org.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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 *16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, Cancun, Mexico, December 18-21, 2017*, pages 364–371. IEEE.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. Dbpedia - A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19274–19286. PMLR.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *ACM Comput. Surv.*, 56(9):230:1–230:39.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8086–8098. Association for Computational Linguistics.

Julian J. McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013*, pages 165–172. ACM.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-supervised hierarchical text classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6826–6833. AAAI Press.

Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *CoRR*, abs/2402.06196.

Lorenzo Paletto, Valerio Basile, and Roberto Esposito. 2024. Label augmentation for zero-shot hierarchical text classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7697–7706. Association for Computational Linguistics.

Silviu Pitis, Michael R. Zhang, Andrew Wang, and Jimmy Ba. 2023. Boosted prompt ensembles for large language models. *CoRR*, abs/2304.05970.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2695–2709. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Fabian Schmidt, Karin Hammerfald, Henrik Haaland Jahren, Amir Hossein Payberah, and Vladimir Vlassov. 2024. Single-pass hierarchical text classification with large language models. In *IEEE International Conference on Big Data, BigData 2024, Washington, DC, USA, December 15-18, 2024*, pages 5412–5421. IEEE.

Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021a. Taxoclass: Hierarchical multi-label text classification using only class names. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4239–4249. Association for Computational Linguistics.

Jianhao Shen, Yichun Yin, Lin Li, Lifeng Shang, Xin Jiang, Ming Zhang, and Qun Liu. 2021b. Generate & rank: A multi-task framework for math word problems. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2269–2279. Association for Computational Linguistics.

Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto. 2018. HFT-CNN: learning hierarchical category structure for multi-label short text categorization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 811–816. Association for Computational Linguistics.

Aixin Sun and Ee-Peng Lim. 2001. Hierarchical text classification and evaluation. In *Proceedings of the 2001 IEEE International Conference on Data Mining, 29 November - 2 December 2001, San Jose, California, USA*, pages 521–528. IEEE Computer Society.

Seyed Amin Tabatabaei, Sarah Fancher, Michael Parsons, and Arian Askari. 2024. Can large language models serve as effective classifiers for hierarchical multi-label classification of scientific documents at industrial scale? *CoRR*, abs/2412.05137.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, and 38 others. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. 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 (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7109–7119. 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3740–3751. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2550–2575. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3912–3921. Association for Computational Linguistics.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024a. Self-contrast: Better reflection through inconsistent solving perspectives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3602–3622. Association for Computational Linguistics.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024b. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1946–1965. Association for Computational Linguistics.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2025. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. In *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pages 2032–2042. ACM.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Prompt consistency for zero-shot task generalization. In *Findings of the Association for Computational Linguistics:*

EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 2613–2626. 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, ACL 2020, Online, July 5-10, 2020*, pages 1106–1117. Association for Computational Linguistics.

A Cost Analysis

In this section, we conduct a cost analysis for HiEPS. The primary cost of HiEPS lies in the input side, and in our experiment, we have adopted the *input token caching* technique to reduce the computational cost by caching the prompting strategies (instructions). This allows the model only to process the given document for each new incoming raw input, which has reduced more than half of our time consumption. Based on this, we report the average running time and cost *per sample* on DBpedia using GPT-3.5 (0.5/1.5 USD per 1M input/output tokens) in Table 7, indicating that:

1. HiEPS achieves superior performance while incurring higher resource consumption, i.e., **x3.3 running time and x2.6 cost compared to the best baseline** SC with pvv (n=10).

2. **Simply increasing the computational overhead of baselines is insufficient to match the performance of HiEPS**, where we upgrade the best baseline SC with pvv by sampling n=40 times, which surpasses the running time and cost of HiEPS, while its performance still falls behind HiEPS by a margin of **1.25%**, **4.73%**, and **2.75%** on the three metrics.

3. HiEPS is a method that only involves LLM inference, with a cost of **less than 1 cent per sample**, making the overall cost modest.

HiEPS with prompting strategy selection. We further introduce *HiEPS-SS*, namely, HiEPS with prompting strategy selection, which selectively discards strategies to reduce resource consumption under the scenario where a validation set is given. During the selection, the strategies are ranked by their contribution to HiEPS from least to most significant, and progressively removed from HiEPS in this order, until its performance drop on Micro-F1 reaches 1% on the validation set. Note that the contribution of each strategy is measured by the performance drop when removing a strategy from HiEPS on the validation set. As shown in Table 8, while maintaining competitive performances (performance drops of less than 1%), HiEPS-SS can significantly reduce the resource consumption, namely, **with a reduction of about 50% computational cost and 25%~50% running time**.

Furthermore, one can use taxonomy pruning (Zhang et al., 2025), or fast inference techniques like speculative decoding (Leviathan et al., 2023) to further reduce the resource consumption of HiEPS. We leave these explorations to future works.

B Additional Experimental Results

B.1 Self-Consistency with Different Base Prompting Strategies

Since SC commonly integrates Chain-of-Thought (CoT) prompting to support step-by-step thinking, it aligns better with global strategies (global-BFS and global-DFS), which are designed to encourage holistic structure understanding and decision-making, rather than with local strategies (per-path, per-parent, and flattened). In fact, global-DFS generally outperforms global-BFS, which is shown in Figure 2. Therefore, we adopt global-DFS as the base strategy for SC (as well as for CoT) in our main experiments. The experimental results in Table 9 further illustrate that global-DFS is the strongest base strategy for SC, where it achieves the best results on two of the three datasets.

In Table 9, we also show the results of other prompting strategies combined with SC using GPT-3.5 on WOS, DBpedia, and Amazon, where we report the results of **w/o SC** (direct prompting), **SC-mv** (SC with majority voting), and **SC-pvv** (SC with path-valid voting). We perform n=40 samplings for SC to ensure that the resource consumption of these baselines is at the same level as that of HiEPS. For per-parent and flattened prompting, only majority voting is adopted, since path-valid voting is not compatible with them. The results indicate that, though applying SC (especially SC-pvv) can improve the performance of direct prompting in most cases[6], *the performance of SC with any single prompting strategy still lags*

[6]In a few cases, SC-mv or SC-pvv slightly underperforms direct prompting, such as when using Global-BFS on WOS and the Macro-F1 on WOS. We speculate that for such knowledge-oriented tasks like WOS, the Chain-of-Thought prompting in SC might interfere with the model's ability to make correct judgments.

Table 7: The average running time and cost (per sample) of different methods on DBpedia using GPT-3.5.

Method	Time (s)	Input Tokens	Output Tokens	Cost (1e-3 USD)	Micro-F1	Macro-F1	Example-F1
Per-path	0.8	1805	7	0.91	80.43	70.00	79.70
Per-parent	1.8	1111	6	0.56	78.40	64.10	76.41
Flattened	0.7	816	3	0.41	82.06	75.67	78.93
Global-BFS	0.8	1902	12	0.97	74.54	66.34	73.48
Global-DFS	1.0	2088	22	1.08	81.02	71.64	79.57
Per-path-multi	1.1	1826	8	0.93	61.68	55.99	67.43
Per-parent-muti	2.0	1134	6	0.58	63.43	61.27	65.65
Flattened-multi	1.0	837	4	0.42	69.02	66.08	71.58
CoT	1.8	2106	94	1.19	82.20	73.68	80.51
SC with pvv (n=10)	2.8	2106	813	2.27	88.72	77.44	86.43
ToT	2.1	1413	20	0.74	81.50	69.57	80.06
HiEPS	9.1	11518	67	5.86	**91.87**	**84.61**	**91.87**
SC with pvv (n=40)	**13.9**	2106	3294	**5.99**	90.62	79.88	89.12

Table 8: Results of HiEPS with prompting strategy selection (HiEPS-SS). #St. denotes the number of the adopted strategies. Time ↓ and Cost ↓ indicate the reduction ratio of running time and cost, respectively. Experiments are conducted using GPT-3.5.

Dataset	Method	#St.	Micro-F1	Macro-F1	Example-F1	Time ↓	Cost ↓	Discarded Strategies
WOS	HiEPS	8	65.41	55.11	65.39	-	-	-
	HiEPS-SS	4	65.16	54.19	65.11	50%	54%	Per-parent, Per-path, Global-BFS, Global-DFS
DBpedia	HiEPS	8	91.87	84.61	91.87	-	-	-
	HiEPS-SS	5	90.97	84.34	90.97	30%	48%	Per-path-multi, Per-path, Global-BFS
Amazon	HiEPS	6	67.07	29.99	67.01	-	-	-
	HiEPS-SS	4	66.18	29.42	66.11	25%	54%	Per-path-multi, Per-path

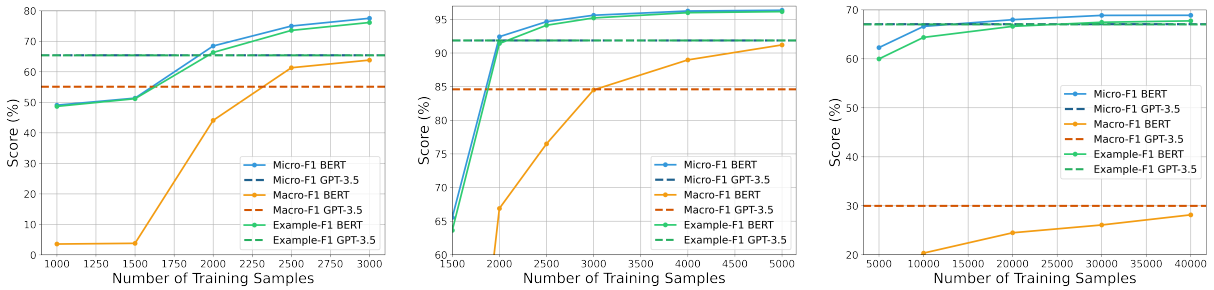
Table 9: Comparisons of different prompting strategies combined with Self-Consistency using GPT-3.5. The best result is bold and the second best is underlined. ↑ means the improvement of HiEPS over the best SC baselines.

Dataset	Method	Micro-F1			Macro-F1			Example-F1		
		w/o SC	SC-mv	SC-pvv	w/o SC	SC-mv	SC-pvv	w/o SC	SC-mv	SC-pvv
WOS	Per-path	59.77	57.80	60.51	50.53	48.75	49.24	58.62	57.72	60.25
	Per-parent	57.21	59.15	-	45.52	48.43	-	56.14	57.95	-
	Flattened	58.65	59.14	-	47.23	46.27	-	55.02	58.38	-
	Global-BFS	57.65	56.46	56.46	48.09	46.79	44.21	56.63	56.44	56.41
	Global-DFS	57.79	58.89	58.74	49.02	45.80	44.19	56.79	58.61	58.60
	HiEPS	**65.41** ↑4.9			**55.11** ↑4.6			**65.39** ↑5.1		
DBpedia	Per-path	80.43	81.52	84.27	70.00	70.00	73.69	79.70	80.63	82.44
	Per-parent	78.40	82.98	-	64.10	69.51	-	76.41	81.69	-
	Flattened	82.06	83.90	-	75.67	76.72	-	78.93	80.44	-
	Global-BFS	74.54	79.56	87.92	66.34	70.30	76.98	73.48	78.64	86.00
	Global-DFS	81.02	85.65	90.62	71.64	77.25	79.88	79.57	84.52	89.12
	HiEPS	**91.87** ↑1.3			**84.61** ↑4.7			**91.87** ↑2.8		
Amazon	Per-path	60.37	61.69	64.36	24.73	25.59	26.44	59.96	61.63	63.78
	Per-parent	62.30	62.61	-	25.86	27.08	-	61.91	61.70	-
	Global-BFS	59.98	59.30	60.65	22.50	23.61	24.56	59.49	59.23	59.48
	Global-DFS	60.90	66.01	66.08	24.11	27.76	27.85	60.39	65.30	66.05
	HiEPS	**67.07** ↑1.0			**29.99** ↑2.1			**67.01** ↑1.0		

behind HiEPS, which highlights that a simple ensemble over multiple decoding paths with a single prompting strategy is insufficient to tackle the core challenge of HTC, namely, understanding the hierarchical structure and making decisions accurately. In contrast, HiEPS achieves superior performance by integrating the strengths of different prompting strategies, enabling the model to comprehend the hierarchical structure from multiple perspectives and effectively integrate their decisions.

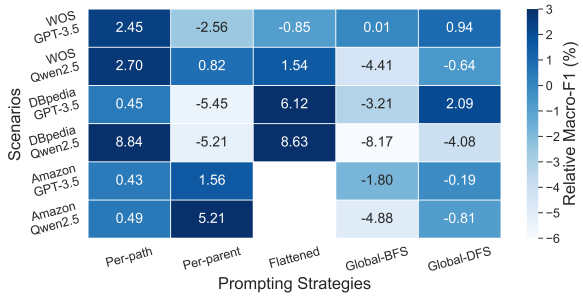
B.2 Heat Maps of Macro-F1 and Example-F1

Following the analysis of relative Micro-F1 scores in Figure 2, we further present the heat maps il-

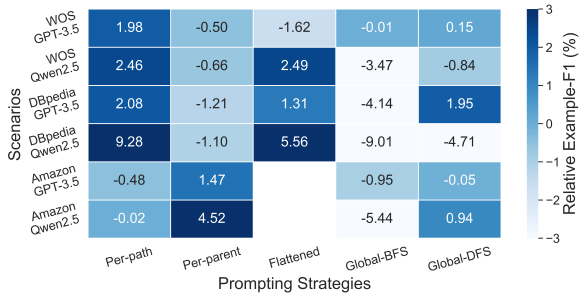


(a) Comparison on WOS. (b) Comparison on DBpedia. (c) Comparison on Amazon.

Figure 9: Comparison with BERT trained on datasets of different sizes. The results of HiEPS are depicted with dashed lines, whereas those of BERT are shown with solid lines.



(a) Comparison on relative Macro-F1.



(b) Comparison on relative Example-F1.

Figure 10: The relative Macro-F1 and Example-F1 of different prompting strategies under different scenarios

lustrating the relative Macro-F1 and Example-F1 scores of different prompting strategies, in Figure 10. We observe that the differences among strategies become even more substantial in terms of Macro-F1. Overall, these results reaffirm the key observations that: (i) within a single scenario, different prompting strategies result in notably diverse performance; and (ii) across different scenarios, the effectiveness of each strategy varies considerably. This highlights the instability of relying on a single prompting strategy for HTC with LLMs.

B.3 Comparisons with Supervised Method

We further compare HiEPS with the supervised method to provide an intuitive illustration of the current capabilities of LLMs in HTC. In particular, we fine-tune BERT (bert-base-uncased) on datasets of different sizes, using a binary cross-entropy loss over the entire label space. We use an NVIDIA RTX A5000 GPU to train the model for 50 epochs with Adam optimizer with a learning rate of $3e-5$. The batch size is fixed as 32 with a maximum sequence length of 512.

As shown in Figure 9, on the WOS dataset, HiEPS attains comparable Example-F1 and Micro-F1 scores to those of BERT trained with 2,000 labeled instances, while surpassing it in Macro-F1. On DBpedia, HiEPS achieves Example-F1 and Micro-F1 results similar to BERT trained with 2,000 instances and matches the Macro-F1 performance of BERT trained with 3,000 instances. For the Amazon dataset, HiEPS performs on par with BERT trained with 10,000 to 20,000 instances in Example-F1 and Micro-F1 and notably outperforms BERT trained with 40,000 samples in Macro-F1. These results suggest that, with the support of HiEPS, LLMs can effectively substitute supervised methods under limited annotation budgets. Furthermore, LLMs exhibit a particular strength in handling long-tail categories, a prevalent challenge in HTC, thus yielding superior performance in Macro-F1 compared to supervised counterparts.

C More Related Works

Hierarchical Text Classification methods based on deep learning (Kowsari et al., 2017) can be broadly categorized into local and global approaches, where local approaches use deep neural networks as local classifiers to classify child nodes (Kowsari et al., 2017; Shimura et al., 2018; Banerjee et al., 2019), while global approaches focus on building a global classifier for the entire hierarchy, such as using recursive regularization (Gopal and Yang, 2013), Graph Neural Network (GNN)-based

encoder (Huang et al., 2019), or joint document label embedding space (Chen et al., 2021). Recent methods based on pre-trained language models (PLM) like BERT have demonstrated further performance gains through prompt-tuning (Wang et al., 2022b) or contrastive learning (Wang et al., 2022a). However, these methods require large amounts of labeled data, which is often expensive and time-consuming to collect. To address this, recent works have explored HTC under limited resources, including weakly-supervised HTC (Meng et al., 2019; Shen et al., 2021a; Zhang et al., 2025) that only requires unlabeled training data, few-shot HTC (Ji et al., 2023; Chen et al., 2024; Ji et al., 2024) which requires a few labeled data, and zero-shot HTC (Bhambhoria et al., 2023; Bongiovanni et al., 2023; Paletto et al., 2024) which does not require any training samples or labels. These methods leverage the prior knowledge within PLMs to perform HTC via similarity matching, entailment prediction, and particularly, prompting large language models (Minaee et al., 2024).

D More Discussions

Discussions on In-Context Learning. To enhance LLM's reliability, another commonly used method is to conduct in-context learning (Brown et al., 2020), which leverages the model's ability to learn from examples provided in the prompt. However, on the one hand, this technique requires collecting few-shot labeled data, which is particularly difficult for rare classes. On the other hand, in-context learning is not as effective for HTC due to the extreme semantic ambiguity in the expansive hierarchical label sets (Chen et al., 2024). Chen et al. (2024) address this issue through elaborately trained encoders that promote the effectiveness of similarity matching on fine-grained labels when retrieving few-shot examples. In general, the investigation of in-context learning falls beyond the scope of our study.

Distinction between HTC and Reasoning Tasks. Moreover, we distinguish our proposed method from reasoning-based approaches. While both HTC and reasoning are complex tasks, their core challenges differ fundamentally: reasoning focuses on multi-step logical deduction, whereas HTC involves comprehending the semantics and relationships within the large and structural taxonomies. To address this, HiEPS is specifically designed for HTC, which leverages diverse perspectives to interpret the label hierarchy and integrate their respective advantages. Empirically, we observe that advanced reasoning methods like Chain-of-Thought prompting (Wei et al., 2022) are less effective for HTC tasks that are knowledge-oriented, like WOS, where HiEPS achieves notably stronger results.

E Full Prompt Design

The full prompt designs of different prompting strategies are listed in Table 10 to 17.

Table 10: Full prompts of **per-path** prompting on WOS.

You will be provided with an academic abstract, and please **select its domain type from the following taxonomy paths**: [**civil engineering, green building; medical science, birth control; electrical engineering, digital control; psychology, problem-solving; psychology, schizophrenia; medical science, hepatitis c; biochemistry, enzymology; psychology, prejudice; medical science, fungal infection; computer science, parallel computing; civil engineering, smart material; mechanical engineering, manufacturing engineering; mechanical engineering, thermodynamics; civil engineering, water pollution; mechanical engineering, computer-aided design; electrical engineering, state space representation; medical science, psoriasis; psychology, gender roles; medical science, dementia; psychology, depression; medical science, alzheimer's disease; civil engineering, ambient intelligence; medical science, migraine; psychology, prenatal development; biochemistry, polymerase chain reaction; medical science, polycythemia vera; computer science, symbolic computation; medical science, psoriatic arthritis; computer science, cryptography; electrical engineering, pid controller; psychology, antisocial personality disorder; computer science, bioinformatics; medical science, senior health; computer science, operating systems; psychology, borderline personality disorder; medical science, stress management; computer science, algorithm design; medical science, hereditary angioedema; medical science, rheumatoid arthritis; psychology, false memories; psychology, attention; medical science, hypothyroidism; medical science, parkinson's disease; psychology, prosocial behavior; medical science, digestive health; psychology, media violence; medical science, headache; medical science, idiopathic pulmonary fibrosis; electrical engineering, system identification; medical science, atopic dermatitis; medical science, mental health; medical science, hiv/aids; electrical engineering, electrical generator; medical science, menopause; civil engineering, remote sensing; biochemistry, cell biology; medical science, emergency contraception; computer science, data structures; computer science, distributed computing; electrical engineering, electrical circuits; psychology, seasonal affective disorder; medical science, cancer; civil engineering, stealth technology; electrical engineering, operational amplifier; biochemistry, southern blotting; medical science, medicare; medical science, overactive bladder; mechanical engineering, machine design; medical science, sports injuries; electrical engineering, signal-flow graph; biochemistry, human metabolism; medical science, irritable bowel syndrome; computer science, structured storage; electrical engineering, lorentz force law; medical science, anxiety; medical science, sprains and strains; computer science, computer programming; biochemistry, immunology; mechanical engineering, hydraulics; computer science, machine learning; medical science, osteoarthritis; electrical engineering, microcontroller; psychology, child abuse; medical science, myelofibrosis; psychology, eating disorders; medical science, children's health; medical science, skin care; computer science, software engineering; medical science, osteoporosis; electrical engineering, electric motor; mechanical engineering, fluid mechanics; medical science, multiple sclerosis; computer science, image processing; civil engineering, suspension bridge; medical science, allergies; civil engineering, solar energy; computer science, network security; biochemistry, genetics; civil engineering, construction management; electrical engineering, control engineering; medical science, smoking cessation; civil engineering, rainwater harvesting; medical science, addiction; computer science, computer graphics; medical science, autism; psychology, social cognition; biochemistry, northern blotting; computer science, computer vision; mechanical engineering, internal combustion engine; electrical engineering, voltage law; electrical engineering, electrical network; biochemistry, molecular biology; psychology, person perception; mechanical engineering, materials engineering; medical science, crohn's disease; medical science, kidney health; medical science, weight loss; medical science, heart disease; medical science, lymphoma; electrical engineering, electricity; medical science, asthma; psychology, leadership; medical science, ankylosing spondylitis; medical science, low testosterone; electrical engineering, analog signal processing; computer science, relational databases; medical science, atrial fibrillation; medical science, bipolar disorder; medical science, diabetes; mechanical engineering, strength of materials; medical science, parenting; medical science, healthy sleep; civil engineering, geotextile; psychology, nonverbal communication**]. Just give the taxonomy path as shown in the provided list.

Table 11: Full prompts of **per-parent** prompting on WOS.

Step 1
You will be provided with an academic abstract, and please **select its domain type from the following categories**: [**mechanical engineering, medical science, electrical engineering, computer science, civil engineering, biochemistry, psychology**]. Just give the category names as shown in the provided list. Each of these candidate categories contains a set of fine-grained sub-categories as follows: mechanical engineering: (fluid mechanics, hydraulics, computer-aided design, manufacturing engineering, machine design, thermodynamics, materials engineering, strength of materials, internal combustion engine); medical science: (alzheimer's disease, parkinson's disease, sprains and strains, cancer, sports injuries, senior health, multiple sclerosis, hepatitis c, weight loss, low testosterone, fungal infection, diabetes, parenting, birth control, heart disease, allergies, menopause, emergency contraception, skin care, myelofibrosis, hypothyroidism, headache, overactive bladder, irritable bowel syndrome, polycythemia vera, atrial fibrillation, smoking cessation, lymphoma, asthma, bipolar disorder, crohn's disease, idiopathic pulmonary fibrosis, mental health, dementia, rheumatoid arthritis, osteoporosis, medicare, psoriatic arthritis, addiction, atopic dermatitis, digestive health, healthy sleep, anxiety, psoriasis, ankylosing spondylitis, children's health, stress management, hiv/aids, migraine, osteoarthritis, hereditary angioedema, kidney health, autism); electrical engineering: (electric motor, digital control, microcontroller, electrical network, electrical generator, electricity, operational amplifier, analog signal processing, state space representation, signal-flow graph, electrical circuits, lorentz force law, system identification, pid controller, voltage law, control engineering); computer science: (symbolic computation, computer vision, computer graphics, operating systems, machine learning, data structures, network security, image processing, parallel computing, distributed computing, algorithm design, computer programming, relational databases, software engineering, bioinformatics, cryptography, structured storage); civil engineering: (green building, water pollution, smart material, ambient intelligence, construction management, suspension bridge, geotextile, stealth technology, solar energy, remote sensing, rainwater harvesting); biochemistry: (molecular biology, enzymology, southern blotting, northern blotting, human metabolism, polymerase chain reaction, immunology, genetics, cell biology); psychology: (prenatal development, attention, eating disorders, borderline personality disorder, prosocial behavior, false memories, problem-solving, prejudice, antisocial personality disorder, nonverbal communication, leadership, child abuse, gender roles, depression, social cognition, seasonal affective disorder, person perception, media violence, schizophrenia).
Step 2
You will be provided with an academic abstract, and please **select its domain type from the following categories**: [**symbolic computation, computer vision, computer graphics, operating systems, machine learning, data structures, network security, image processing, parallel computing, distributed computing, algorithm design, computer programming, relational databases, software engineering, bioinformatics, cryptography, structured storage**]. Just give the category names as shown in the provided list. These candidate categories belong to the same coarse-grained category: computer science.

Table 12: Full prompts of **flattened** prompting on WOS.

You will be provided with an academic abstract, and please **select its domain type from the following categories**: [**green building, birth control, digital control, problem-solving, schizophrenia, hepatitis c, enzymology, prejudice, fungal infection, parallel computing, smart material, manufacturing engineering, thermodynamics, water pollution, computer-aided design, state space representation, psoriasis, gender roles, dementia, depression, alzheimer's disease, ambient intelligence, migraine, prenatal development, polymerase chain reaction, polycythemia vera, symbolic computation, psoriatic arthritis, cryptography, pid controller, antisocial personality disorder, bioinformatics, senior health, operating systems, borderline personality disorder, stress management, algorithm design, hereditary angioedema, rheumatoid arthritis, false memories, attention, hypothyroidism, parkinson's disease, prosocial behavior, digestive health, media violence, headache, idiopathic pulmonary fibrosis, system identification, atopic dermatitis, mental health, hiv/aids, electrical generator, menopause, remote sensing, cell biology, emergency contraception, data structures, distributed computing, electrical circuits, seasonal affective disorder, cancer, stealth technology, operational amplifier, southern blotting, medicare, overactive bladder, machine design, sports injuries, signal-flow graph, human metabolism, irritable bowel syndrome, structured storage, lorentz force law, anxiety, sprains and strains, computer programming, immunology, hydraulics, machine learning, osteoarthritis, microcontroller, child abuse, myelofibrosis, eating disorders, children's health, skin care, software engineering, osteoporosis, electric motor, fluid mechanics, multiple sclerosis, image processing, suspension bridge, allergies, solar energy, network security, genetics, construction management, control engineering, smoking cessation, rainwater harvesting, addiction, computer graphics, autism, social cognition, northern blotting, computer vision, internal combustion engine, voltage law, electrical network, molecular biology, person perception, materials engineering, crohn's disease, kidney health, weight loss, heart disease, lymphoma, electricity, asthma, leadership, ankylosing spondylitis, low testosterone, analog signal processing, relational databases, atrial fibrillation, bipolar disorder, diabetes, strength of materials, parenting, healthy sleep, geotextile, nonverbal communication**]. Just give the category names as shown in the provided list.

Table 13: Full prompts of **gloabl-BFS** prompting on WOS.

You are a helpful assistant for the task of academic abstract classification on the Web Of Science dataset. This dataset has a hierarchical labeling structure with two levels of labels:

Domains (Level 1):

[**mechanical engineering, medical science, electrical engineering, computer science, civil engineering, biochemistry, psychology**]

Subdomains (Level 2):

[**fluid mechanics, hydraulics, computer-aided design, manufacturing engineering, machine design, thermodynamics, materials engineering, strength of materials, internal combustion engine**] (subdomains of mechanical engineering); [**alzheimer's disease, parkinson's disease, sprains and strains, cancer, sports injuries, senior health, multiple sclerosis, hepatitis c, weight loss, low testosterone, fungal infection, diabetes, parenting, birth control, heart disease, allergies, menopause, emergency contraception, skin care, myelofibrosis, hypothyroidism, headache, overactive bladder, irritable bowel syndrome, polycythemia vera, atrial fibrillation, smoking cessation, lymphoma, asthma, bipolar disorder, crohn's disease, idiopathic pulmonary fibrosis, mental health, dementia, rheumatoid arthritis, osteoporosis, medicare, psoriatic arthritis, addiction, atopic dermatitis, digestive health, healthy sleep, anxiety, psoriasis, ankylosing spondylitis, children's health, stress management, hiv/aids, migraine, osteoarthritis, hereditary angioedema, kidney health, autism**] (subdomains of medical science); [**electric motor, digital control, microcontroller, electrical network, electrical generator, electricity, operational amplifier, analog signal processing, state space representation, signal-flow graph, electrical circuits, lorentz force law, system identification, pid controller, voltage law, control engineering**] (subdomains of electrical engineering); [**symbolic computation, computer vision, computer graphics, operating systems, machine learning, data structures, network security, image processing, parallel computing, distributed computing, algorithm design, computer programming, relational databases, software engineering, bioinformatics, cryptography, structured storage**] (subdomains of computer science); [**green building, water pollution, smart material, ambient intelligence, construction management, suspension bridge, geotextile, stealth technology, solar energy, remote sensing, rainwater harvesting**] (subdomains of civil engineering); [**molecular biology, enzymology, southern blotting, northern blotting, human metabolism, polymerase chain reaction, immunology, genetics, cell biology**] (subdomains of biochemistry); [**prenatal development, attention, eating disorders, borderline personality disorder, prosocial behavior, false memories, problem-solving, prejudice, antisocial personality disorder, nonverbal communication, leadership, child abuse, gender roles, depression, social cognition, seasonal affective disorder, person perception, media violence, schizophrenia**] (subdomains of psychology);

You will be provided with an academic abstract, and please **select its domain type and subdomain type from the above label hierarchy**. Separate the selected domain and subdomain by '||'. Just give the category names as shown in the provided label hierarchy.

Table 14: Full prompts of **gloabl-DFS** prompting on WOS.

You are a helpful assistant for the task of academic abstract classification on the Web Of Science dataset. This dataset consists of seven domains, each containing multiple subdomains. The domain label structure is as follows:

"computer science": ["symbolic computation", "computer vision", "computer graphics", "operating systems", "machine learning", "data structures", "network security", "image processing", "parallel computing", "distributed computing", "algorithm design", "computer programming", "relational databases", "software engineering", "bioinformatics", "cryptography", "structured storage"], "medical science": ["alzheimer's disease", "parkinson's disease", "sprains and strains", "cancer", "sports injuries", "senior health", "multiple sclerosis", "hepatitis c", "weight loss", "low testosterone", "fungal infection", "diabetes", "parenting", "birth control", "heart disease", "allergies", "menopause", "emergency contraception", "skin care", "myelofibrosis", "hypothyroidism", "headache", "overactive bladder", "irritable bowel syndrome", "polycythemia vera", "atrial fibrillation", "smoking cessation", "lymphoma", "asthma", "bipolar disorder", "crohn's disease", "idiopathic pulmonary fibrosis", "mental health", "dementia", "rheumatoid arthritis", "osteoporosis", "medicare", "psoriatic arthritis", "addiction", "atopic dermatitis", "digestive health", "healthy sleep", "anxiety", "psoriasis", "ankylosing spondylitis", "children's health", "stress management", "hiv/aids", "migraine", "osteoarthritis", "hereditary angioedema", "kidney health", "autism"], "civil engineering": ["green building", "water pollution", "smart material", "ambient intelligence", "construction management", "suspension bridge", "geotextile", "stealth technology", "solar energy", "remote sensing", "rainwater harvesting"], "electrical engineering": ["electric motor", "digital control", "microcontroller", "electrical network", "electrical generator", "electricity", "operational amplifier", "analog signal processing", "state space representation", "signal-flow graph", "electrical circuits", "lorentz force law", "system identification", "pid controller", "voltage law", "control engineering"], "biochemistry": ["molecular biology", "enzymology", "southern blotting", "northern blotting", "human metabolism", "polymerase chain reaction", "immunology", "genetics", "cell biology"], "mechanical engineering": ["fluid mechanics", "hydraulics", "computer-aided design", "manufacturing engineering", "machine design", "thermodynamics", "materials engineering", "strength of materials", "internal combustion engine"], "psychology": ["prenatal development", "attention", "eating disorders", "borderline personality disorder", "prosocial behavior", "false memories", "problem-solving", "prejudice", "antisocial personality disorder", "nonverbal communication", "leadership", "child abuse", "gender roles", "depression", "social cognition", "seasonal affective disorder", "person perception", "media violence", "schizophrenia"]

You will be provided with an academic abstract, and please **select its domain type and subdomain type from the above label hierarchy**. Separate the selected domain and subdomain by '||'. Just give the category names as shown in the provided label structure.

Table 15: Full prompts of **per-path-multi** prompting on WOS.

You will be provided with an academic abstract, and please **select its all possible domain types from the following taxonomy paths**: [**civil engineering, green building; medical science, birth control; electrical engineering, digital control; psychology, problem-solving; psychology, schizophrenia; medical science, hepatitis c; biochemistry, enzymology; psychology, prejudice; medical science, fungal infection; computer science, parallel computing; civil engineering, smart material; mechanical engineering, manufacturing engineering; mechanical engineering, thermodynamics; civil engineering, water pollution; mechanical engineering, computer-aided design; electrical engineering, state space representation; medical science, psoriasis; psychology, gender roles; medical science, dementia; psychology, depression; medical science, alzheimer's disease; civil engineering, ambient intelligence; medical science, migraine; psychology, prenatal development; biochemistry, polymerase chain reaction; medical science, polycythemia vera; computer science, symbolic computation; medical science, psoriatic arthritis; computer science, cryptography; electrical engineering, pid controller; psychology, antisocial personality disorder; computer science, bioinformatics; medical science, senior health; computer science, operating systems; psychology, borderline personality disorder; medical science, stress management; computer science, algorithm design; medical science, hereditary angioedema; medical science, rheumatoid arthritis; psychology, false memories; psychology, attention; medical science, hypothyroidism; medical science, parkinson's disease; psychology, prosocial behavior; medical science, digestive health; psychology, media violence; medical science, headache; medical science, idiopathic pulmonary fibrosis; electrical engineering, system identification; medical science, atopic dermatitis; medical science, mental health; medical science, hiv/aids; electrical engineering, electrical generator; medical science, menopause; civil engineering, remote sensing; biochemistry, cell biology; medical science, emergency contraception; computer science, data structures; computer science, distributed computing; electrical engineering, electrical circuits; psychology, seasonal affective disorder; medical science, cancer; civil engineering, stealth technology; electrical engineering, operational amplifier; biochemistry, southern blotting; medical science, medicare; medical science, overactive bladder; mechanical engineering, machine design; medical science, sports injuries; electrical engineering, signal-flow graph; biochemistry, human metabolism; medical science, irritable bowel syndrome; computer science, structured storage; electrical engineering, lorentz force law; medical science, anxiety; medical science, sprains and strains; computer science, computer programming; biochemistry, immunology; mechanical engineering, hydraulics; computer science, machine learning; medical science, osteoarthritis; electrical engineering, microcontroller; psychology, child abuse; medical science, myelofibrosis; psychology, eating disorders; medical science, children's health; medical science, skin care; computer science, software engineering; medical science, osteoporosis; electrical engineering, electric motor; mechanical engineering, fluid mechanics; medical science, multiple sclerosis; computer science, image processing; civil engineering, suspension bridge; medical science, allergies; civil engineering, solar energy; computer science, network security; biochemistry, genetics; civil engineering, construction management; electrical engineering, control engineering; medical science, smoking cessation; civil engineering, rainwater harvesting; medical science, addiction; computer science, computer graphics; medical science, autism; psychology, social cognition; biochemistry, northern blotting; computer science, computer vision; mechanical engineering, internal combustion engine; electrical engineering, voltage law; electrical engineering, electrical network; biochemistry, molecular biology; psychology, person perception; mechanical engineering, materials engineering; medical science, crohn's disease; medical science, kidney health; medical science, weight loss; medical science, heart disease; medical science, lymphoma; electrical engineering, electricity; medical science, asthma; psychology, leadership; medical science, ankylosing spondylitis; medical science, low testosterone; electrical engineering, analog signal processing; computer science, relational databases; medical science, atrial fibrillation; medical science, bipolar disorder; medical science, diabetes; mechanical engineering, strength of materials; medical science, parenting; medical science, healthy sleep; civil engineering, geotextile; psychology, nonverbal communication**]. Separate your choices by '||'. Just give the taxonomy paths as shown in the provided list.

Table 16: Full prompts of **per-parent-multi** prompting on WOS.

Step 1
You will be provided with an academic abstract, and please **select its all possible domain types from the following categories**: [**mechanical engineering, medical science, electrical engineering, computer science, civil engineering, biochemistry, psychology**]. Separate your choices by '||'. Just give the category names as shown in the provided list. Each of these candidate categories contains a set of fine-grained sub-categories as follows: mechanical engineering: (fluid mechanics, hydraulics, computer-aided design, manufacturing engineering, machine design, thermodynamics, materials engineering, strength of materials, internal combustion engine); medical science: (alzheimer's disease, parkinson's disease, sprains and strains, cancer, sports injuries, senior health, multiple sclerosis, hepatitis c, weight loss, low testosterone, fungal infection, diabetes, parenting, birth control, heart disease, allergies, menopause, emergency contraception, skin care, myelofibrosis, hypothyroidism, headache, overactive bladder, irritable bowel syndrome, polycythemia vera, atrial fibrillation, smoking cessation, lymphoma, asthma, bipolar disorder, crohn's disease, idiopathic pulmonary fibrosis, mental health, dementia, rheumatoid arthritis, osteoporosis, medicare, psoriatic arthritis, addiction, atopic dermatitis, digestive health, healthy sleep, anxiety, psoriasis, ankylosing spondylitis, children's health, stress management, hiv/aids, migraine, osteoarthritis, hereditary angioedema, kidney health, autism); electrical engineering: (electric motor, digital control, microcontroller, electrical network, electrical generator, electricity, operational amplifier, analog signal processing, state space representation, signal-flow graph, electrical circuits, lorentz force law, system identification, pid controller, voltage law, control engineering); computer science: (symbolic computation, computer vision, computer graphics, operating systems, machine learning, data structures, network security, image processing, parallel computing, distributed computing, algorithm design, computer programming, relational databases, software engineering, bioinformatics, cryptography, structured storage); civil engineering: (green building, water pollution, smart material, ambient intelligence, construction management, suspension bridge, geotextile, stealth technology, solar energy, remote sensing, rainwater harvesting); biochemistry: (molecular biology, enzymology, southern blotting, northern blotting, human metabolism, polymerase chain reaction, immunology, genetics, cell biology); psychology: (prenatal development, attention, eating disorders, borderline personality disorder, prosocial behavior, false memories, problem-solving, prejudice, antisocial personality disorder, nonverbal communication, leadership, child abuse, gender roles, depression, social cognition, seasonal affective disorder, person perception, media violence, schizophrenia).
Step 2
You will be provided with an academic abstract, and please **select its all possible domain types from the following taxonomy paths**: [**computer science, symbolic computation; computer science, computer vision; computer science, computer graphics; computer science, operating systems; computer science, machine learning; computer science, data structures; computer science, network security; computer science, image processing; computer science, parallel computing; computer science, distributed computing; computer science, algorithm design; computer science, computer programming; computer science, relational databases; computer science, software engineering; computer science, bioinformatics; computer science, cryptography; computer science, structured storage; biochemistry, molecular biology; biochemistry, enzymology; biochemistry, southern blotting; biochemistry, northern blotting; biochemistry, human metabolism; biochemistry, polymerase chain reaction; biochemistry, immunology; biochemistry, genetics; biochemistry, cell biology**]. Separate your choices by '||'. Just give the taxonomy paths as shown in the provided list.

Table 17: Full prompts of **flattened-multi** prompting on WOS.

You will be provided with an academic abstract, and please **select its all possible domain types from the following categories**: [**green building, birth control, digital control, problem-solving, schizophrenia, hepatitis c, enzymology, prejudice, fungal infection, parallel computing, smart material, manufacturing engineering, thermodynamics, water pollution, computer-aided design, state space representation, psoriasis, gender roles, dementia, depression, alzheimer's disease, ambient intelligence, migraine, prenatal development, polymerase chain reaction, polycythemia vera, symbolic computation, psoriatic arthritis, cryptography, pid controller, antisocial personality disorder, bioinformatics, senior health, operating systems, borderline personality disorder, stress management, algorithm design, hereditary angioedema, rheumatoid arthritis, false memories, attention, hypothyroidism, parkinson's disease, prosocial behavior, digestive health, media violence, headache, idiopathic pulmonary fibrosis, system identification, atopic dermatitis, mental health, hiv/aids, electrical generator, menopause, remote sensing, cell biology, emergency contraception, data structures, distributed computing, electrical circuits, seasonal affective disorder, cancer, stealth technology, operational amplifier, southern blotting, medicare, overactive bladder, machine design, sports injuries, signal-flow graph, human metabolism, irritable bowel syndrome, structured storage, lorentz force law, anxiety, sprains and strains, computer programming, immunology, hydraulics, machine learning, osteoarthritis, microcontroller, child abuse, myelofibrosis, eating disorders, children's health, skin care, software engineering, osteoporosis, electric motor, fluid mechanics, multiple sclerosis, image processing, suspension bridge, allergies, solar energy, network security, genetics, construction management, control engineering, smoking cessation, rainwater harvesting, addiction, computer graphics, autism, social cognition, northern blotting, computer vision, internal combustion engine, voltage law, electrical network, molecular biology, person perception, materials engineering, crohn's disease, kidney health, weight loss, heart disease, lymphoma, electricity, asthma, leadership, ankylosing spondylitis, low testosterone, analog signal processing, relational databases, atrial fibrillation, bipolar disorder, diabetes, strength of materials, parenting, healthy sleep, geotextile, nonverbal communication**]. Separate your choices by '||'. Just give the category names as shown in the provided list.