MKT: A Multi-Stage Knowledge Transfer Framework to Mitigate Catastrophic Forgetting in Multi-Domain Chinese Spelling Correction

Peng Xing[1*], **Yinghui Li**[1*], **Shirong Ma**[1], **Xinnian Liang**[2], **Haojing Huang**[1]
Yangning Li[1], **Shu-Yu Guo**[1], **Hai-Tao Zheng**[1†], **Wenhao Jiang**[3†], **Ying Shen**[4]
[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] State Key Lab of Software Development Environment, Beihang University
[3] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
[4] School of Intelligent Systems Engineering, Sun-Yat Sen University
{xp23, liyinghu20}@mails.tsinghua.edu.cn

Abstract

Chinese Spelling Correction (CSC) aims to detect and correct spelling errors in given sentences. Recently, multi-domain CSC has gradually attracted the attention of researchers because it is more practicable. In this paper, we focus on the key flaw of the CSC model when adapting to multi-domain scenarios: the tendency to forget previously acquired knowledge upon learning new domain-specific knowledge (i.e., **catastrophic forgetting**). To address this, we propose a novel model-agnostic **M**ulti-stage **K**nowledge **T**ransfer (**MKT**) framework with an evolving teacher model and dynamic distillation weights for knowledge transfer in each domain, rather than focusing solely on new domain knowledge. It deserves to be mentioned that we are the first to apply continual learning methods to the multi-domain CSC task. Experiments prove our method's effectiveness over traditional approaches, highlighting the importance of overcoming catastrophic forgetting to enhance model performance.

1 Introduction

Chinese Spelling Correction (CSC) plays a critical role in detecting and correcting spelling errors in Chinese text (Li et al., 2022c; Ma et al., 2022), enhancing the accuracy of technologies like Optical Character Recognition (OCR) and Automatic Speech Recognition (ASR) (Afli et al., 2016; Wang et al., 2018). In search engines, for example, CSC reduces human error, ensuring that users find the information they seek accurately.

In real applications, the input text may come from various domains, demanding that the model contains different domain-specific knowledge. As illustrated in figure 1, the word "强基(Strong Foundation)" is evidently common in the Chinese Education domain. Accurately correcting "张(open)" to "强(Strong)" requires the model to have specific knowledge about the Chinese Education domain. Therefore, some works have begun to focus on the impact of domain-specific knowledge on the performance of CSC models (Lv et al., 2023a; Wu et al., 2023). These types of knowledge are difficult to grasp through the original model's generalization ability. However, with the rise of the internet and social media, a large number of new internet slang, codes, and colloquialisms emerge every year, traditional static training paradigms struggle to meet these demands.

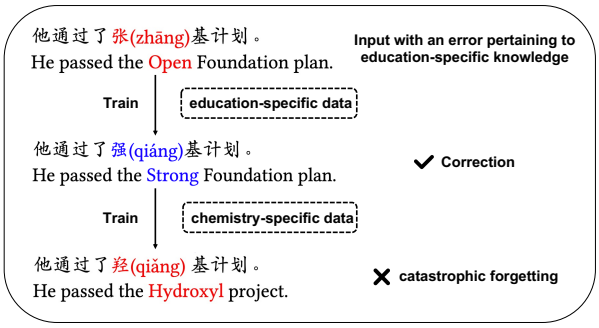

Figure 1: Case of model forgetting general-domain knowledge during continual learning. red represents the misspelled character and blue represents the corrected character.

To dynamically integrate newly acquired knowledge, two main paradigms exist. The first method entails periodically blending old data with new and retraining the model to refresh its knowledge base. However, this technique requires significant computational resources, as it involves repeatedly reprocessing data that has already been encountered. The second paradigm, continual learning, addresses this challenge by updating the model incrementally, eliminating the need for full retraining on past data. This significantly improves resource efficiency and stands as a key research direction for developing models that can continuously adapt to an ever-evolving linguistic landscape.

[*] indicates equal contribution.
[†] Corresponding authors (cswhjiang@gmail.com, zheng.haitao@sz.tsinghua.edu.cn).

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 451–462
November 4-9, 2025

The core challenge of the continual learning setting is to minimize catastrophic forgetting of previously acquired knowledge while learning in new domains (Wang et al., 2024). As demonstrated in Figure 1, when a CSC model learns educational-specific knowledge, it accurately corrects the word "强基(Strong Foundation)". However, after it continues to learn knowledge from the chemistry domain, it would learn the new knowledge of "羟基(hydroxyl)", but forget the education word "强基(Strong Foundation)". *Unfortunately, in previous multi-domain CSC studies, the challenge of this catastrophic forgetting of domain-specific knowledge has not been fully explored.*

As a widely adopted method in continual learning, knowledge distillation leverages a static teacher model and incorporates task-specific heads to handle diverse tasks. However, this approach has two significant limitations: firstly, the static teacher model struggles to adapt to the ever-expanding knowledge from emerging domains; secondly, the continuous addition of task-specific heads leads to an increasingly bloated model. When directly applied to CSC tasks, these knowledge distillation methods inevitably encounter the same scalability challenges. Given the wide variety of error types in Chinese text, the model grows progressively more complex and computationally demanding, rendering it impractical for real-world applications.

To address these limitations, we propose a novel, model-agnostic Multi-stage Knowledge Transfer (MKT) framework. Unlike traditional methods that require expanding the model architecture, the MKT framework maintains a fixed structure, avoiding model expansion. This not only reduces computational overhead but also simplifies deployment in real-world environments, where linguistic data is dynamic and constantly evolving.

MKT framework incorporates two dynamic mechanisms to enhance continual learning. First, the evolving teacher model adapts to growing domain-specific knowledge, overcoming the limitations of static teacher models. Second, it seamlessly integrates knowledge from both new and existing domains by transferring its accumulated knowledge to the current student model, with distillation weights dynamically adjusted based on the data ratio. The combination of these two dynamic mechanisms makes MKT framework significantly more effective than other continual learning methods, greatly mitigating catastrophic forgetting.

2 Related Work

2.1 Chinese Spelling Correction

In the field of CSC, we witness significant advancements in various model architectures and modules, as evidenced by recent works (Li et al., 2022b, 2023b; Zhang et al., 2023; Ye et al., 2023b, 2022; Ma et al., 2023; Ye et al., 2023a; Huang et al., 2023; Li et al., 2023d). Early models such as the Confusionset-guided Pointer Networks focus on optimizing at the dataset level by leveraging confusion sets for character generation. This technique enhances accuracy by considering commonly confused characters (Wang et al., 2019). Innovations in embeddings, like the REALISE model, improve model inputs by integrating semantic, phonetic, and visual information into character embeddings, thereby enriching the representational capacity of the model (Xu et al., 2021). Improvements in encoders are highlighted by models such as Soft-Masked BERT, which employs Soft MASK techniques post-detection to blend input characters with [MASK] embeddings. This method is effective for error prediction and has shown significant improvements in performance (Zhang et al., 2020). Another notable model, SpellGCN, constructs a character graph and maps it to interdependent detection classifiers based on BERT-extracted representations, showcasing innovative uses of graph neural networks in spelling correction (Cheng et al., 2020).

Previous research in multi-domain CSC emphasizes cross-domain knowledge sharing and generalization (Lv et al., 2023a). Typically, this involves training models on high-quality datasets to generalize effectively to specific domains. However, domain-specific knowledge is hard to generalize, and fine-tuning on multiple datasets can lead to catastrophic forgetting, where new knowledge overwrites old knowledge. This paper addresses catastrophic forgetting by introducing mechanisms that balance retaining existing knowledge with integrating new information. We propose a framework that mitigates forgetting while ensuring robust performance across multiple domains.

2.2 Continual Learning

In the field of continual learning, core strategies such as replay, regularization, and parameter isolation play pivotal roles (Liu et al., 2022; Li et al., 2022a; Wang et al., 2023; Dong et al., 2023; Li et al., 2023c).Replay methods, including techniques like GEM and MER, work by retaining train-

ing samples and using constraints or meta-learning to align gradients effectively (Lopez-Paz and Ranzato, 2017; Riemer et al., 2018). Regularization strategies, such as Elastic Weight Consolidation (EWC), aim to preserve task-specific knowledge by assigning higher importance to parameters crucial for previous tasks (Kirkpatrick et al., 2017). However, both our initial investigations and existing literature (Buzzega et al., 2020) suggest that knowledge distillation techniques, such as Learning without Forgetting (LwF), tend to outperform EWC. Knowledge distillation is a method that enables incremental training by transferring knowledge from larger models to smaller ones, thereby facilitating the integration of new knowledge while preserving previously learned information (Gou et al., 2021). Parameter isolation techniques, such as CL-plugin, address task interference by assigning dedicated parameters to different tasks, thus minimizing the risk of overlap and interference (Ke et al., 2022). However, CL-plugin is specifically designed for task-incremental learning, it is not suitable for our domain-incremental learning experiments.

Our MKT framework stands out as a model-agnostic approach, capable of being applied across various CSC models. By leveraging the strengths of existing continual learning strategies and integrating them into a cohesive framework, we aim to effectively mitigate catastrophic forgetting and enhance the adaptability of CSC models in multi-domain scenarios.

3 Our Approach

3.1 Problem Formulation

The CSC task is to detect and correct spelling errors in Chinese texts. Given a misspelled sentence $X = \{x_1, x_2, ..., x_m\}$ with m characters, a CSC model takes X as input, detects possible spelling errors at character level, and outputs a corresponding correct sentence $Y = \{y_1, y_2, ..., y_m\}$ of equal length. This task can be viewed as a conditional sequence generation problem that models the probability of $p(Y|X)$. In multi-domain CSC tasks under a continual learning setting, assuming that there are n domains $D = \{D_1, D_2, ..., D_n\}$, these domains are trained sequentially, where each domain D_k is trained without access to the data from previous domains, from D_1 to D_{k-1}. Furthermore, after training domain D_k, we should consider the performance of all domains from D_1 to D_k, a metric which we will introduce in Section 4.1.

3.2 Structure of MKT Framework

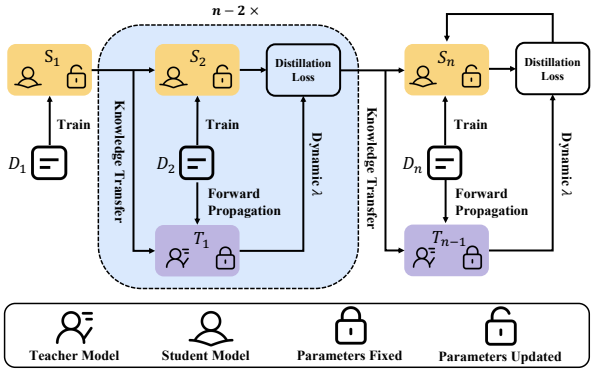

Figure 2: Overview of the MKT framework and the pipeline for multi-domain training.

As illustrated in Figure 2, teacher model acts as a comprehensive knowledge repository, effectively serving as a backup of the student model from the previous stage to calculate the distillation loss for the current stage's student model. It encapsulates all the domain-specific knowledge accumulated to date, providing crucial guidance for the model training in the current phase. This dynamically evolving teacher model avoids the substantial cost of continually updating a static teacher's knowledge repository in traditional knowledge distillation and prevents any expansion of the teacher model. Additionally, we conduct experiments to explore how to a priori select appropriate distillation weights (results are shown in Table 3), so that framework can dynamically adjust distillation weights during training to achieve better performance.

3.3 MKT Framework for Multi-domain CSC

We consider the scenario where the training is comprised of n stages, denoted by $k = 1, 2..., n$. At k-th stage, a subset of data $\{x_k^{(i)}, y_k^{(i)}\}_{i=1}^{G_k}$ are fed to the model, where G_k refers to the number of samples at k-th domain, $x_k^{(i)}$ refers to i-th sample at k-th domain. Assume that $u_k(\cdot)$ is an target function that maps each $x_k^{(i)}$ to $y_k^{(i)}$ at stage k, i.e., $y_k^{(i)} = u_k(x_k^{(i)})$. Under the continual learning setting, our goal is to train a CSC model $g(\cdot\,; w)$ parameterized by w, such that $g(\cdot\,; w)$ not only fits well to $u_k(\cdot)$, but also fits $u_{k-1}(\cdot)\,, u_{k-2}(\cdot), \cdots, u_1(\cdot)$ in early stages to alleviate catastrophic forgetting. We optimize model parameters by minimizing the loss function:

$$L^{(k)} = \lambda L_s^{(k)} + L_h^{(k)}. \quad (1)$$

In the equation, λ is a hyper-parameter that

ranges from $[0, 1]$. $L_s^{(k)}$ is the knowledge distillation loss, calculating cross entropy between the output probabilities of teacher model $g(\cdot\,; w_{k-1})$ and student model $g(\cdot\,; w_k)$:

$$L_s^{(k)} = -\sum_{i=1}^{T_k} g(x_k^{(i)};\, \omega_{k-1}) \times \log g(x_k^{(i)};\, \omega_k). \tag{2}$$

$L_h^{(k)}$ is the cross-entropy loss between the output of student model $g(\cdot\,; w_k)$ and ground truth y_k:

$$L_h^{(k)} = -\sum_{i=1}^{T_k} y_k^{(i)} \times \log g(x_k^{(i)};\, \omega_k). \tag{3}$$

The choice of λ is related to the ratio of domain data and old data, S_d is domain data scale and S_o is old data scale:

$$\lambda = \frac{S_d}{S_o} \tag{4}$$

Algorithm 1 MKT Framework

Input: Training set D_k, Student model S_{k-1}
Output: Student model S_k

1: Copy S_{k-1} as the teacher model T_k
2: Freeze the parameters of T_k
3: Calculate λ according to Equation 4.
4: S_k forward propagation and calculates the loss guided by T_k according to Equation 1
5: Optimize the parameters of S_k
6: **Return** S_k

As shown in Algorithm 1, during the training of the k-th domain, both the teacher model T_{k-1} and the student model S_{k-1} are initialized using the parameters of S_{k-1}. The teacher model T_{k-1} is kept frozen to provide stable knowledge guidance, while the student model S_{k-1} is further optimized on the current domain data, resulting in the updated model S_k. The final loss is the dynamically weighted summation of the knowledge distillation loss $L_s^{(k)}$ and the original CSC task loss $L_h^{(k)}$, with the weights as shown in Equation 4.

4 Experiment and Result

4.1 Datasets and Metrics

For domain selection, we defined four domains: **General, Car, Medical, and Legal**. As shown in Table 1, CSC models trained on the general dataset exhibit significant knowledge gaps compared to later domains. This closely aligns with our focus on catastrophic forgetting of domain-specific knowledge, making these datasets and experiments highly suitable for our research. As shown in Table 1, the Zero-Shot performance of CSC-specific models trained on the General dataset exhibits significant knowledge gaps compared to later domains. This closely aligns with our focus on catastrophic forgetting of domain-specific knowledge, making these datasets and experiments highly suitable for our research. For the General domain, we also use SIGHAN13/14/15 (Wu et al., 2013; Yu and Li, 2014; Tseng et al., 2015) and Wang271K (Wang et al., 2018) as training data and SIGHAN15 test set as our test data. For other special domains, we utilize the data resources released by LEMON (Wu et al., 2023) and ECSpell (Lv et al., 2023b), and randomly take 500 samples from the original data of each domain as the test set. The dataset statistics are presented in the Table A.

Our evaluation predominantly relies on the sentence-level F1 score, a widely acknowledged metric (Xu et al., 2021).

In each table, Avg represents the overall performance after training on all domains. Unlike average accuracy (AA) (Wang et al., 2023), we use the average sentence-level F1 score, which is a more stringent metric than AA.

4.2 Baseline Methods

To validate the model-agnostic nature of MKT, we select three commonly used CSC models with different architectures as baselines, aiming to evaluate our approach across various frameworks. As shown in Table 1, these include the ***RoBERTa*** (Liu et al., 2019), ***Soft-Masked-BERT*** (Zhang et al., 2020), and ***REALISE*** (Xu et al., 2021), which integrates multimodal information. In addition, we evaluate the Chinese spelling correction capabilities of two advanced general large language models (LLMs): ***LLama-3.1-8B*** (Dubey et al., 2024) and ***Qwen-2.5-7B*** (Yang et al., 2024). However, their performance shows a significant gap compared to specialized CSC models, so we conduct our experiments on different continual learning methods using these specialized CSC models.

To validate the effectiveness of our MKT framework, we compare it with different continual learning methods on the aforementioned models, thereby demonstrating the superiority of our approach. Specifically, we conduct experiments with ***Fine-tuning*** (lacking

Type	Model	Method	General	CAR	MED	LAW	Avg
CSC	RoBERTa	Zero-Shot	73.40	0.85	19.55	25.84	29.91
		Fine-tuning	67.41	33.50	42.86	62.35	51.53
	Soft-Masked BERT	Zero-Shot	69.60	15.53	30.53	46.84	40.63
		Fine-tuning	54.22	30.73	43.88	68.54	49.34
	REALISE	Zero-Shot	77.84	0.28	19.45	28.57	31.54
		Fine-tuning	70.78	27.48	53.33	70.59	55.55
General	LLama-3.1-8B	Zero-Shot	8.48	3.80	13.62	14.73	10.16
		Fine-tuning	9.34	9.76	18.15	36.98	18.56
	Qwen-2.5-7B	Zero-Shot	20.66	14.00	28.00	48.30	27.74
		Fine-tuning	22.22	20.16	30.60	64.75	34.43

Table 1: Baseline Performance of General LLMs and Specialized Models on Each Domain's Test Set.

Model	Method	General	CAR	MED	LAW	Avg
RoBERTa	Upper Bound	73.40	39.14	45.13	62.88	55.14
	Replay(random)	70.07	34.87	41.33	59.51	51.45
	Replay(RAP)	70.09	36.22	43.00	58.25	51.89
	EWC	67.77	35.64	40.09	62.75	51.56
	MKT(Ours)	68.58	36.18	43.56	62.47	**52.70**$^{\uparrow}$
Soft-Masked BERT	Upper Bound	69.60	45.78	58.28	70.68	61.09
	Replay(random)	47.45	23.88	39.51	64.30	43.79
	Replay(RAP)	54.48	22.86	46.52	60.59	46.11
	EWC	54.00	25.72	45.21	69.65	48.65
	MKT(Ours)	60.90	35.64	52.21	70.40	**54.79**$^{\uparrow}$
REALISE	Upper Bound	77.84	32.82	56.62	70.85	59.53
	Replay(random)	75.78	27.83	53.81	69.25	56.67
	Replay(RAP)	76.10	31.51	50.33	69.76	56.93
	EWC	74.11	28.22	54.22	70.16	56.68
	MKT(Ours)	73.84	31.25	54.1	70.18	**57.34**$^{\uparrow}$

Table 2: Final Performance After Continual Learning Across All Domains.

any forgetting-prevention mechanisms), two replay-based methods—***random sampling*** and ***RAP*** (**R**eplay **A**ccording im**P**ortance) and a regularization-based approach (***EWC***). We also include ***knowledge distillation*** methods, with the corresponding experiments discussed in our ablation study. Additionally, we provide implementation details in the appendix B.

4.3 Results and Analyses

Main Results From Table 2, it can be seen that after applying the MKT framework, whether it is RoBERTa, Soft-Masked BERT specially designed for CSC, or REALISE that integrates multi-modal information, their performance in all domains improves compared to Fine-Tuning without any forgetting-prevention strategy. This fully demonstrates the advantages of our proposed MKT framework in terms of both effectiveness and model-agnostic capability. When comparing MKT with other continual learning methods, among the two replay strategies, RAP outperforms random replay and even approaches the performance of MKT. However, it requires ten times the training data of MKT, leading to significantly higher training overhead. EWC can also effectively mitigate catastrophic forgetting, but it still lags behind MKT, underscoring the superiority of MKT among various continual learning approaches.

Parameter Study To explore the impact of the key parameter λ, we conduct experiments using REALISE + MKT on the General dataset and three subsequent specific domain datasets, selecting a portion of the General dataset as the old dataset. As shown in Table 3, the old dataset size is set to 50, 20, and 10 times that of the corresponding

$\frac{S_d}{S_o}$	λ	General	MED	Avg	General	MED	Avg	General	LAW	Avg
0.02	0	66.73	30.42	48.58	65.65	47.17	56.41	66.07	59.45	62.76
	0.01	66.97	30.25	48.61	65.78	47.60	56.69	66.9	58.27	62.59
	0.02	67.26	30.96	**49.11**$^{\uparrow}$	66.67	47.27	**56.97**$^{\uparrow}$	66.91	59.01	**62.96**$^{\uparrow}$
	0.04	67.51	30.19	48.85	67.17	45.16	56.17	66.84	58.38	62.61
0.05	0	62.25	29.85	46.05	62.36	42.44	52.40	61.27	58.5	59.89
	0.025	64.10	29.42	46.76	61.84	43.89	52.87	61.87	59.84	60.86
	0.05	65.80	30.17	**47.99**$^{\uparrow}$	63.82	41.96	**52.89**$^{\uparrow}$	62.30	60.63	**61.47**$^{\uparrow}$
	0.1	64.85	30.17	47.51	62.94	41.27	52.11	63.00	57.82	60.41
0.1	0	55.07	27.44	41.26	57.69	40.18	48.94	53.68	54.94	54.31
	0.05	58.46	28.03	43.25	57.48	41.91	49.70	54.53	55.62	55.08
	0.1	59.13	28.51	**43.82**$^{\uparrow}$	58.81	41.38	**50.10**$^{\uparrow}$	55.66	55.20	**55.43**$^{\uparrow}$
	0.2	57.47	25.70	41.59	58.96	35.70	47.33	55.60	54.22	54.91

Table 3: Selection of Optimal Distillation Weights (λ) under Different Domain(S_d) and Old(S_o) Data Ratios.

specific domain dataset. When λ is set to 0.5, 1, and 2 times the ratio of the domain dataset size to the old dataset size, the experimental results show stable improvements over the baseline (i.e., $\lambda = 0$). In particular, when λ matches the ratio of domain data to old data, all domains achieve the best performance. Thus, for MKT an appropriate λ can be chosen based on the ratio of new domain data to old data to achieve optimal performance.

Model	Method	Buffer size	General	CAR	MED	LAW	Avg
REALISE	Replay (random)	0.001	74.14	27.56	54.07	67.88	55.91
		0.01	75.78	27.83	53.81	69.25	**56.67**$^{\uparrow}$
		0.1	74.44	30.33	51.94	67.87	56.15
	Replay (RAP)	0.001	74.31	26.77	49.37	67.88	54.58
		0.01	75.48	31.51	48.75	68.67	56.10
		0.1	76.10	31.51	50.33	69.76	**56.93**$^{\uparrow}$

Table 4: Replay Performance with Buffer Sizes.

Buffer study We investigated the optimal buffer sizes for two replay methods. As presented in Table 4, random sampling delivers peak performance with a buffer size equivalent to 1% of the old data, striking an effective balance between new and old data scales. In contrast, importance-based sampling (RAP) excels at incorporating critical knowledge from both domains, achieving optimal results with a buffer size of 10% of the old data, albeit at the cost of extended training time.

Catastrophic Forgetting The above analysis convincingly demonstrate that the MKT framework outperforms other continual learning methods in overall performance after training across all domains. To better observe the forgetting at each stage when training on subsequent domain datasets, we select the best-performing model from Table 2 (i.e., REALISE) and examine its performance

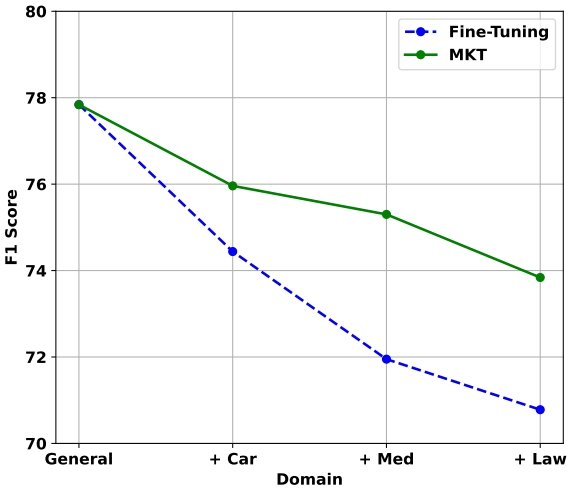

Figure 3: The phenomenon of model forgetting General-domain knowledge during incremental domain training.

loss (i.e., catastrophic forgetting) on the General dataset after incremental training with data from other domains, as shown in Figure 3. The performance loss of REALISE on the General dataset is much smoother when optimized with MKT, indicating that MKT framework effectively mitigates catastrophic forgetting at each stage.

Method	General	CAR	MED	LAW	Avg
KD	74.39	30.03	48.80	64.63	54.46
Replay(RAP)	75.71	30.94	44.42	69.55	55.16
MKT(Ours)	74.07	30.03	51.55	67.13	**55.69**$^{\uparrow}$

Table 5: Average Performance of Two Training Orders.

Training Order In real-world applications, data often arrive in unpredictable, continuously expanding orders. To simulate this uncertainty, we randomly select two training orders and evaluate the best-performing model, REALISE, from Table 2. Table 5 shows the average domain performance for both orders, demonstrating that the MKT framework effectively integrates domain-specific knowledge under varied training sequences and mitigates catastrophic forgetting.

4.4 Ablation Study

MKT differs from knowledge distillation in two key aspects: evolving teacher model and dynamically distillation weights. To evaluate their effectiveness, we conduct ablation experiments on REALISE without these optimizations. Table 6 shows that both mechanisms of MKT effectively mitigate catastrophic forgetting. A continuously evolving

Method	General	CAR	MED	LAW	Avg
KD	74.23	29.69	52.68	67.61	56.05
+ evolving teacher	72.74	29.25	55.28	70.85	57.03
+ dynamic λ	74.13	30.30	53.74	67.34	56.38
MKT(Ours)	73.84	31.25	54.10	70.18	**57.34**$^{\uparrow}$

Table 6: The Impact of Dynamic Distillation Weights (λ) and the Evolving Teacher Model on Performance.

Circumventing Catastrophic Forgetting	
Input	青量级玩乐SUV
+CAR(Fine-tuning)	轻量级玩乐SUV
+CAR(+MKT)	轻量级玩乐SUV
+MED(Fine-tuning)	氰量级玩乐SUV
+MED(+MKT)	轻量级玩乐SUV
Target	轻量级玩乐SUV
Input	百年纪念板排放量:6.0L
+CAR(Fine-tuning)	百年纪念版排放量:6.0L
+CAR(+MKT)	百年纪念版排放量:6.0L
+MED(Fine-tuning)	百年纪念板排放量:6.0L
+MED(+MKT)	百年纪念版排放量:6.0L
Target	百年纪念版排放量:6.0L

Table 7: Cases from the CAR Test Set, Conducted on the REALISE Model, Show that the MKT Framework Mitigates Catastrophic Forgetting.

teacher model can incorporate the most important knowledge previously learned, effectively reducing the student's forgetting of prior knowledge. For dynamic distillation weights, we provided experimental results in Table 3. MKT's adaptation to the ratio of domain and old data allows it to better learn the most important knowledge from domain and old data. Using dynamic distillation weights alone can only provide limited performance improvement, with the performance improvement brought by the dynamically evolving teacher being more significant. By integrating both mechanisms, MKT achieves superior anti-forgetting performance in domain adaptation, despite slightly higher forgetting on the General dataset compared to the fixed teacher method.

5 Conclusion

This paper demonstrates through experimentation that existing CSC models, when adapting to multi-domain scenarios, tend to forget previously acquired domain-specific knowledge, a phenomenon known as catastrophic forgetting. To address this, we propose an effective, model-agnostic MKT framework that incorporates an evolving teacher model and dynamic distillation weights. This framework balances retaining existing knowledge with integrating new information, effectively mitigating catastrophic forgetting. Extensive experiments and detailed analyses underscore the importance of tackling catastrophic forgetting, proving that our approach outperforms other continual learning approaches.

6 Case Study

To further verify the effectiveness of our MKT in mitigating catastrophic forgetting in multi-domain CSC, we present some cases in Table 7. For a test sentence in the CAR domain, REALISE accurately corrects errors after fine-tuning on CAR. However, after further fine-tuning on the MED domain, it can no longer correct successfully and instead predicts "氰(cyanide)" related to the medical domain. Another case of catastrophic forgetting is overcorrection, such as when the character "版(version)" is mistakenly corrected to "板(board)" after learning from the MED domain. Both cases illustrate classic examples of catastrophic forgetting where old domain knowledge is washed away by new domain knowledge. It can be seen that with the optimization of MKT, REALISE effectively avoids the occurrence of catastrophic forgetting.

7 Limitations

We do not compare our proposed method against commonly used Large Language Models (LLMs) (Kuang et al., 2025; Li et al., 2024b; Huang et al., 2024; Li et al., 2025a; Zhang et al., 2025b; Xu et al., 2025; Yu et al., 2024; Li et al., 2025b,c) in our experiments. The primary reason is that in the CSC task, representative LLMs still lag behind traditional fine-tuned smaller models, which has been proved by many related works (Li et al., 2023a, 2024a, 2025d; Ye et al., 2024b,a, 2025; Zhang et al., 2025a; Zou et al., 2025), which has been confirmed by many related works, and we also verify this in Table 1. In addition, our approach specifically focuses on the Chinese scenarios. However, other languages, such as English, could also benefit from our methodology. We will conduct related studies on English scenarios in the future.

References

Haithem Afli, Zhengwei Qiu, Andy Way, and Páraic Sheridan. 2016. Using SMT for OCR error correction of historical texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 962–966, Portorož, Slovenia. European Language Resources Association (ELRA).

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. 2020. Dark experience for general continual learning: a strong, simple baseline. *ArXiv*, abs/2004.07211.

Xingyi Cheng, Weidi Xu, Kunlong Chen, Shaohua Jiang, Feng Wang, Taifeng Wang, Wei Chu, and Yuan Qi. 2020. SpellGCN: Incorporating phonological and visual similarities into language models for Chinese spelling check. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 871–881, Online. Association for Computational Linguistics.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2023. A survey of natural language generation. *ACM Comput. Surv.*, 55(8):173:1–173:38.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

Haojing Huang, Jingheng Ye, Qingyu Zhou, Yinghui Li, Yangning Li, Feng Zhou, and Hai-Tao Zheng. 2023. A frustratingly easy plug-and-play detection-and-reasoning module for chinese spelling check. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11514–11525. Association for Computational Linguistics.

Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Haitao Zheng. 2024. Lateval: An interactive llms evaluation benchmark with incomplete information from lateral thinking puzzles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10186–10197. ELRA and ICCL.

Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu, and Bing Liu. 2022. Continual training of language models for few-shot learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10216, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025. Natural language understanding and inference with MLLM in visual question answering: A survey. *ACM Comput. Surv.*, 57(8):190:1–190:36.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, and Philip S. Yu. 2025a. Benchmarking multimodal retrieval augmented generation with dynamic VQA dataset and self-adaptive planning agent. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yinghui Li, Haojing Huang, Jiayi Kuang, Yangning Li, Shu-Yu Guo, Chao Qu, Xiaoyu Tan, Hai-Tao Zheng, Ying Shen, and Philip S. Yu. 2025b. Refine knowledge of large language models via adaptive contrastive learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Y. Li, F. Zhou, Haitao Zheng, and Qingyu Zhou. 2023a. On the (in)effectiveness of large language models for chinese text correction. *ArXiv*, abs/2307.09007.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023b. On the (in)effectiveness of large language models for chinese text correction. *CoRR*, abs/2307.09007.

Yinghui Li, Shulin Huang, Xinwei Zhang, Qingyu Zhou, Yangning Li, Ruiyang Liu, Yunbo Cao, Hai-Tao Zheng, and Ying Shen. 2023c. Automatic context pattern generation for entity set expansion. *IEEE Trans. Knowl. Data Eng.*, 35(12):12458–12469.

Yinghui Li, Jiayi Kuang, Haojing Huang, Zhikun Xu, Xinnian Liang, Yi Yu, Wenlian Lu, Yangning Li, Xiaoyu Tan, Chao Qu, Ying Shen, Hai-Tao Zheng, and Philip S. Yu. 2025c. One example shown, many concepts known! counterexample-driven conceptual reasoning in mathematical llms. *CoRR*, abs/2502.10454.

Yinghui Li, Yangning Li, Yuxin He, Tianyu Yu, Ying Shen, and Hai-Tao Zheng. 2022a. Contrastive learning with hard negative entities for entity set expansion. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1077–1086. ACM.

Yinghui Li, Shirong Ma, Shaoshen Chen, Haojing Huang, Shulin Huang, Yangning Li, Hai-Tao Zheng, and Ying Shen. 2025d. Correct like humans: Progressive learning framework for chinese text error correction. *Expert Syst. Appl.*, 265:126039.

Yinghui Li, Shirong Ma, Qingyu Zhou, Zhongli Li, Yangning Li, Shulin Huang, Ruiyang Liu, Chao Li, Yunbo Cao, and Haitao Zheng. 2022b. Learning from the dictionary: Heterogeneous knowledge guided fine-tuning for chinese spell checking. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 238–249. Association for Computational Linguistics.

Yinghui Li, Shang Qin, Jingheng Ye, Shirong Ma, Yangning Li, Libo Qin, Xuming Hu, Wenhao Jiang, Hai-Tao Zheng, and Philip S. Yu. 2024a. Rethinking the roles of large language models in chinese grammatical error correction. *CoRR*, abs/2402.11420.

Yinghui Li, Zishan Xu, Shaoshen Chen, Haojing Huang, Yangning Li, Yong Jiang, Zhongli Li, Qingyu Zhou, Hai-Tao Zheng, and Ying Shen. 2023d. Towards real-world writing assistance: A chinese character checking benchmark with faked and misspelled characters. *CoRR*, abs/2311.11268.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022c. The past mistake is the future wisdom: Error-driven contrastive probability optimization for Chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3202–3213, Dublin, Ireland. Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S. Yu. 2024b. When llms meet cunning texts: A fallacy understanding benchmark for large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ruiyang Liu, Yinghui Li, Linmi Tao, Dun Liang, and Hai-Tao Zheng. 2022. Are we ready for a new paradigm shift? A survey on visual deep MLP. *Patterns*, 3(7):100520.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

David Lopez-Paz and Marc' Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Qi Lv, Ziqiang Cao, Lei Geng, Chunhui Ai, Xu Yan, and Guohong Fu. 2023a. General and domain-adaptive chinese spelling check with error-consistent pretraining. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(5).

Qi Lv, Ziqiang Cao, Lei Geng, Chunhui Ai, Xu Yan, and Guohong Fu. 2023b. General and domain-adaptive chinese spelling check with error-consistent pretraining. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(5):1–18.

Shirong Ma, Yinghui Li, Haojing Huang, Shulin Huang, Yangning Li, Hai-Tao Zheng, and Ying Shen. 2023. Progressive multi-task learning framework for chinese text error correction. *CoRR*, abs/2306.17447.

Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Li Yangning, Ruiyang Liu, Zhongli Li, Yunbo Cao, Haitao Zheng, and Ying Shen. 2022. Linguistic rules-based corpus generation for native Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 576–589, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. 2018. Learning to learn without forgetting by maximizing transfer and minimizing interference. *ArXiv*, abs/1810.11910.

Yuen-Hsien Tseng, Lung-Hao Lee, Li-Ping Chang, and Hsin-Hsi Chen. 2015. Introduction to SIGHAN 2015 bake-off for Chinese spelling check. In *Proceedings of the Eighth SIGHAN Workshop on Chinese Language Processing*, pages 32–37, Beijing, China. Association for Computational Linguistics.

Dingmin Wang, Yan Song, Jing Li, Jialong Han, and Haisong Zhang. 2018. A hybrid approach to automatic corpus generation for Chinese spelling check. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2517–2527, Brussels, Belgium. Association for Computational Linguistics.

Dingmin Wang, Yi Tay, and Li Zhong. 2019. Confusionset-guided pointer networks for Chinese spelling check. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5780–5785, Florence, Italy. Association for Computational Linguistics.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2023. A comprehensive survey of continual learning: Theory, method and application. *ArXiv*, abs/2302.00487.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. 2024. A comprehensive survey of continual learning: Theory, method and application. *Preprint*, arXiv:2302.00487.

Hongqiu Wu, Shaohua Zhang, Yuchen Zhang, and Hai Zhao. 2023. Rethinking masked language modeling for chinese spelling correction. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1:10743–10756.

Shih-Hung Wu, Chao-Lin Liu, and Lung-Hao Lee. 2013. Chinese spelling check evaluation at SIGHAN bake-off 2013. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, pages 35–42, Nagoya, Japan. Asian Federation of Natural Language Processing.

Heng-Da Xu, Zhongli Li, Qingyu Zhou, Chao Li, Zizhen Wang, Yunbo Cao, Heyan Huang, and Xian-Ling Mao. 2021. Read, listen, and see: Leveraging multimodal information helps Chinese spell checking. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 716–728, Online. Association for Computational Linguistics.

Zhikun Xu, Yinghui Li, Ruixue Ding, Xinyu Wang, Boli Chen, Yong Jiang, Haitao Zheng, Wenlian Lu, Pengjun Xie, and Fei Huang. 2025. Let llms take on the latest challenges! A chinese dynamic question answering benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10435–10448. Association for Computational Linguistics.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 24 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Jingheng Ye, Yinghui Li, Yangning Li, and Hai-Tao Zheng. 2023a. Mixedit: Revisiting data augmentation and beyond for grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10161–10175. Association for Computational Linguistics.

Jingheng Ye, Yinghui Li, Shirong Ma, Rui Xie, Wei Wu, and Hai-Tao Zheng. 2022. Focus is what you need for chinese grammatical error correction. *CoRR*, abs/2210.12692.

Jingheng Ye, Yinghui Li, Qingyu Zhou, Yangning Li, Shirong Ma, Hai-Tao Zheng, and Ying Shen. 2023b. CLEME: debiasing multi-reference evaluation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6174–6189. Association for Computational Linguistics.

Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Peng Xing, Zishan Xu, Guo Cheng, and Zhao Wei. 2024a. EXCGEC: A benchmark of edit-wise explainable chinese grammatical error correction. *CoRR*, abs/2407.00924.

Jingheng Ye, Shang Qin, Yinghui Li, Hai-Tao Zheng, Shen Wang, and Qingsong Wen. 2025. Corrections meet explanations: A unified framework for explainable grammatical error correction. *CoRR*, abs/2502.15261.

Jingheng Ye, Zishan Xu, Yinghui Li, Xuxin Cheng, Linlin Song, Qingyu Zhou, Hai-Tao Zheng, Ying Shen, and Xin Su. 2024b. CLEME2.0: towards more interpretable evaluation by disentangling edits for grammatical error correction. *CoRR*, abs/2407.00934.

Junjie Yu and Zhenghua Li. 2014. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 220–223.

Tianyu Yu, Chengyue Jiang, Chao Lou, Shen Huang, Xiaobin Wang, Wei Liu, Jiong Cai, Yangning Li, Yinghui Li, Kewei Tu, Hai-Tao Zheng, Ningyu Zhang, Pengjun Xie, Fei Huang, and Yong Jiang. 2024. Seqgpt: An out-of-the-box large language model for open domain sequence understanding. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19458–19467. AAAI Press.

Ding Zhang, Yangning Li, Lichen Bai, Hao Zhang, Yinghui Li, Haiye Lin, Hai-Tao Zheng, Xin Su, and Zifei Shan. 2025a. Loss-aware curriculum learning for chinese grammatical error correction. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pages 1–5. IEEE.

Ding Zhang, Yinghui Li, Qingyu Zhou, Shirong Ma, Yangning Li, Yunbo Cao, and Hai-Tao Zheng. 2023. Contextual similarity is more valuable than character similarity: An empirical study for chinese spell checking. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Shaohua Zhang, Haoran Huang, Jicong Liu, and Hang Li. 2020. Spelling error correction with soft-masked BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 882–890, Online. Association for Computational Linguistics.

Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, Peilin Zhou, Aiwei Liu, Yinghui Li, Hao Chen, Jianling Wang, Yu Wang, Feiran Huang,

Sheng Zhou, Jiajun Bu, Allen Lin, James Caverlee, Fakhri Karray, Irwin King, and Philip S. Yu. 2025b. Cold-start recommendation towards the era of large language models (llms): A comprehensive survey and roadmap. *CoRR*, abs/2501.01945.

Deqing Zou, Jingheng Ye, Yulu Liu, Yu Wu, Zishan Xu, Yinghui Li, Hai-Tao Zheng, Bingxu An, Zhao Wei, and Yong Xu. 2025. Revisiting classification taxonomy for grammatical errors. *CoRR*, abs/2502.11890.

A Datasets

Training Set	Domain	Sent	Avg.Length	Errors
Wang271K	General	271,329	42.6	381,962
SIGHAN13	General	700	41.8	343
SIGHAN14	General	3,437	49.6	5,122
SIGHAN15	General	2,338	31.1	3,037
CAR	CAR	2,743	43.4	1,628
MED	MED	3,000	50.2	2,260
LAW	LAW	1,960	30.7	1,681
Test Set	**Domain**	**Sent**	**Avg.Length**	**Errors**
SIGHAN15	General	1,100	30.6	703
CAR	CAR	500	43.7	281
MED	MED	500	49.6	356
LAW	LAW	500	29.7	390

Table 8: Statistics of the Datasets We Use.

B Implementation Details

In the main experiment, we initially train the models on General dataset, which consists of Wang271K combined with double the amount of SIGHAN data. This is followed by training on the CAR, MED, and LAW datasets using various continual learning methods, including Joint-Training, fine-tuning, replay (random), and replay (RAP). Upon completion of training, we evaluate the performance of the final model across all domain-specific datasets to gauge its effectiveness.

Additionally, auxiliary experiments are conducted using our top-performing REALISE model. These experiments investigate several factors such as determining the optimal λ, assessing the appropriate buffer size, examining the effects of different training orders, and performing ablation studies to understand the contribution of each component.

For all experiments, we train the aforementioned datasets for 10 epochs with a batch size of 64. The learning rates are 5e-5 for REALISE and BERT models, and 1e-4 for the Soft-Masked BERT model. Our approach incorporates a knowledge transfer process at each domain, where the λ between L_h and L_s is updated prior to training each domain according to Equation 4. The hyper parameter settings for the auxiliary experiments remain consistent with those used in our main experiments.