CL^{2}GEC: A Multi-Discipline Benchmark for Continual Learning in Chinese Literature Grammatical Error Correction

Shang Qin[1*], **Jingheng Ye**[1*], **Yinghui Li**[1†], **Hai-Tao Zheng**[1,2†], **Qi Li**[3], **Jinxiao Shan**[3], **Zhixing Li**[4], **Hong-Gee Kim** [5],
[1]Shenzhen International Graduate School, Tsinghua University,
[2]Peng Cheng Laboratory, [3]China Merchants Group
[4]Zhipu AI, [5]Seoul National University

Abstract

The growing demand for automated writing assistance in diverse academic domains highlights the need for robust Chinese Grammatical Error Correction (CGEC) systems that can adapt across disciplines. However, existing CGEC research largely lacks dedicated benchmarks for multi-disciplinary academic writing, overlooking continual learning (CL) as a promising solution to handle domain-specific linguistic variation and prevent catastrophic forgetting. To fill this crucial gap, we introduce **CL^{2}GEC**, the first Continual Learning benchmark for Chinese Literature Grammatical Error Correction, designed to evaluate adaptive CGEC across multiple academic fields. Our benchmark includes 10,000 human-annotated sentences spanning 10 disciplines, each exhibiting distinct linguistic styles and error patterns. CL^{2}GEC focuses on evaluating grammatical error correction in a continual learning setting, simulating sequential exposure to diverse academic disciplines to reflect real-world editorial dynamics. We evaluate large language models under sequential tuning, parameter-efficient adaptation, and four representative CL algorithms, using both standard GEC metrics and continual learning metrics adapted to task-level variation. Experimental results reveal that regularization-based methods mitigate forgetting more effectively than replay-based or naive sequential approaches. Our benchmark provides a rigorous foundation for future research in adaptive grammatical error correction across diverse academic domains.[1]

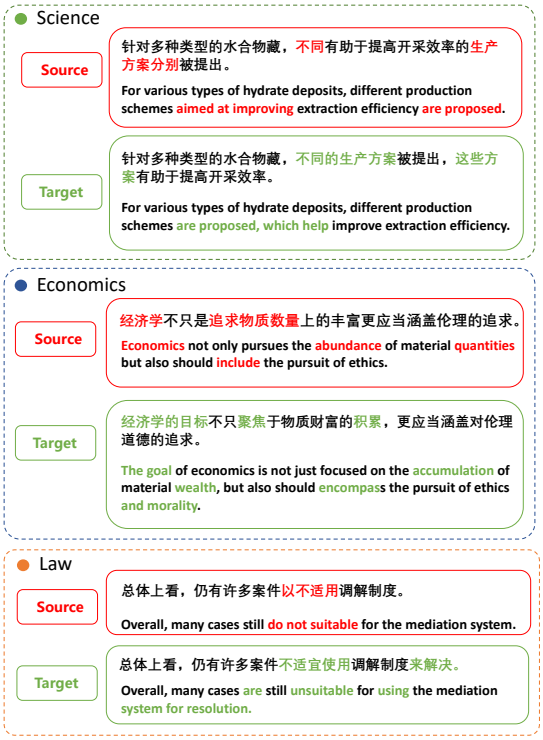

Figure 1: The correction examples in CL^{2}GEC.

1 Introduction

Chinese Grammatical Error Correction (CGEC) has evolved rapidly alongside the surge of large language models (LLMs) (Ye et al., 2025b; Qingsong et al., 2025) and intelligent writing assistants (Li et al., 2022; Qiu et al., 2025; Li et al., 2024b; Zhang et al., 2025). Most existing CGEC benchmarks, however, are (1) learner or general domain oriented (Zhang et al., 2022; Ma et al., 2022), and (2) evaluated in a static setting (Xu et al., 2022; Ye et al., 2023b, 2024). As a result, they offer limited insight into how CGEC models behave in high-stakes professional domains, especially in scientific manuscripts where style, terminology, and error distribution vary markedly across disciplines.

We argue that real-world scientific writing introduces an under-explored challenge for CGEC: *continual domain adaptation* (Wu et al., 2024; Guan et al., 2025). In practice, CGEC systems are expected to ingest papers from, e.g., Physics this

[*]indicates equal contribution.
[†]Corresponding authors: liyinghuihhh@gmail.com, zheng.haitao@sz.tsinghua.edu.cn
[1]Our code and data are publicly available at https://github.com/THUKElab/CL2GEC.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33500–33521
July 2-7, 2026

month and Humanities next month, continually refining its internal knowledge without access to all past data. The threat of catastrophic forgetting (Li et al., 2024a), widely studied in vision (Shmelkov et al., 2017) and NLP tasks (Shao and Feng, 2022), has received almost no attention in CGEC, leaving an open question: *Can modern LLMs retain grammatical knowledge while sequentially adapting to new scientific disciplines?* Figure 1 presents representative correction examples from CL^{2}GEC, illustrating the diversity of grammatical errors encountered in scientific writing.

Addressing this question is crucial for three reasons. First, academia encompasses hundreds of sub-fields whose linguistic conventions differ in syntax and terminology, significantly challenging current LLMs. Second, annotation budgets are usually fragmented by discipline, making one-shot and full-data retraining impractical. Third, reliable cross-domain grammars underpin downstream tasks such as automatic reviewing (Pang et al., 2025), plagiarism detection (Quidwai et al., 2023), and literature summarization (Li et al., 2024d).

Therefore, to systematically study CGEC under the context of continual learning (CL), we present **CL^{2}GEC**, the first Continual Learning benchmark for Chinese Literature Grammatical Error Correction. The CL^{2}GEC benchmark contains 10,000 sentences evenly sampled from 10 academic disciplines, each sentence paired with up to three independent human references. The corpus was curated from China National Knowledge Infrastructure (CNKI)[2], cleaned for copyright, and double-checked by professional editors to reflect authentic error patterns. We release both a canonical split (train/dev/test) and a sequence of 10 task partitions that simulate chronological arrival of disciplines, enabling controlled CL evaluation. CL^{2}GEC aims to set a new standard for evaluating and advancing lifelong grammatical error correction in the era of domain-diverse scientific communication.

Our proposed CL^{2}GEC allows researchers to probe a spectrum of model abilities, mainly including (1) in-domain grammatical accuracy, (2) cross-discipline transfer, and (3) resistance to catastrophic forgetting. Consequently, the benchmark fills a vital gap between generic CGEC datasets and real-world academic editing.

Empirically, we benchmark several representative continual learning strategies, including naive sequential fine-tuning, LoRA-based adaptation (Hu et al., 2021), and four CL algorithms (EWC (Kirkpatrick et al., 2017), GEM (Lopez-Paz and Ranzato, 2017), LwF (Li and Hoiem, 2016), OGD (Farajtabar et al., 2019)). Our comprehensive experiments reveal that while these methods significantly mitigate catastrophic forgetting compared to naive sequential approaches, the optimal strategies vary. We observe a nuanced impact of task ordering on knowledge retention and transfer, including unexpected interference between semantically related disciplines and a trade-off between precision and recall depending on the task sequence.

We highlight our main contributions as follows:

- We introduce CL^{2}GEC, the first large-scale and multi-discipline benchmark tailored to Chinese literature grammatical error correction in the context of continual learning.
- We devise task-specific CL metrics (Average Performance, Backward Transfer) adapted to CGEC, and provide a reproducible evaluation suite.
- We conduct extensive LLM experiments, revealing critical limitations of existing CL methods and establishing solid baselines for future work.

2 Related Work

2.1 Chinese Grammatical Error Correction

Chinese grammatical error correction (CGEC) has developed from early sequence-to-sequence (Seq2Seq) models (Ye et al., 2023a, 2025a; Li et al., 2025), which model correction as a generation task. These approaches benefited from pretraining and syntactic priors but were mainly applied to general learner corpora.

With the advent of large language models (LLMs) (Kuang et al., 2025; Li et al., 2024c), recent work has explored their capabilities for CGEC (Wang et al., 2024; Xiao et al., 2024). Closed- and open-source LLMs have been evaluated through in-context learning and instruction tuning, showing improved fluency and generalization across error types. ScholarGEC (Kong and Wang, 2025) further investigates controllability in academic writing by combining error detection and correction within a multi-task framework.

While prior work focuses on enhancing model performance under static settings, relatively little

[2] https://www.cnki.net/

attention has been paid to domain transfer or continual adaptation. Our CL^{2}GEC benchmark addresses this gap by providing a multi-discipline scientific dataset and evaluating models under domain-incremental settings, enabling more systematic assessment of generalization in CGEC.

2.2 Continual Learning for NLP

Continual learning (CL) aims to enable models to learn from sequentially arriving tasks while mitigating catastrophic forgetting (De Lange et al., 2021; Xing et al., 2024). Existing approaches are typically grouped into three categories: regularization-based, replay-based, and architectural-based.

Regularization-based methods introduce constraints on parameter updates to preserve knowledge of prior tasks. For example, Orthogonal Gradient Descent (OGD) (Farajtabar et al., 2019) enforces gradient orthogonality, Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) penalizes changes to parameters deemed important, and Gradient Episodic Memory (GEM) (Lopez-Paz and Ranzato, 2017) maintains an episodic buffer to avoid loss increases on earlier tasks. Learning without Forgetting (LwF) (Li and Hoiem, 2016) mitigates forgetting through distillation from previous model snapshots.

Replay-based methods explicitly revisit past knowledge by reusing stored samples or their approximations. The simplest variant is Experience Replay, which replays examples from a memory buffer, while more advanced approaches use generative replay or replay-based distillation such as SEEKR (He et al., 2024).

Architectural-based methods adapt the model structure to integrate new information, either by allocating task-specific components or expanding capacity. A representative example is Progressive Prompts (Razdaibiedina et al., 2023), which incrementally composes task-specific prompts into a composite representation.

3 CL^{2}GEC Benchmark

3.1 Problem Definition

Grammatical Error Correction (GEC) aims to transform an ungrammatical sentence $X = \{x_1, x_2, \ldots, x_T\}$ into its grammatically correct counterpart $Y = \{y_1, y_2, \ldots, y_{T'}\}$ while preserving the original semantics. Typically formulated as a sequence-to-sequence task, GEC models are trained to minimize the negative log-likelihood of the corrected output:

$$\mathcal{L}_{\text{GEC}} = -\sum_{t=1}^{T'} \log P(y_t \mid Y_{<t}, X). \quad (1)$$

Continual Learning (CL) addresses the challenge of learning from a stream of tasks $\{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$ without access to previous task data. In the supervised setting, each task $\mathcal{D}_t = \{(x_i^t, y_i^t)\}_{i=1}^{n_t}$ is presented sequentially, and the model is trained to accumulate knowledge over time while avoiding catastrophic forgetting. Let f_Θ be a model with parameters Θ. The goal of CL is to optimize performance across all tasks:

$$\max_{\Theta} \sum_{t=1}^{N} \sum_{(x,y)\in\mathcal{D}_t} \log P_\Theta(y \mid x). \quad (2)$$

Evaluation in CL involves metrics such as *Average Performance* and *Backward Transfer*, which measure the model's ability to retain and transfer knowledge across tasks.

The CL^{2}GEC Task We define CL^{2}GEC as a domain-specific Grammatical Error Correction (GEC) benchmark formulated under the continual learning (CL) paradigm. The dataset is composed of grammatically erroneous academic sentences collected from 10 distinct disciplines (e.g., law, medicine, philosophy), with each domain corresponding to a sequential task:

$$\mathcal{D}_t = \{(X_i^t, Y_i^t)\}_{i=1}^{n_t}. \quad (3)$$

Each pair consists of an erroneous sentence X_i^t and its corrected version Y_i^t. Unlike typical GEC tasks trained on all data simultaneously, CL^{2}GEC is formulated as a continual learning benchmark where the model learns sequentially from each domain and only replays a small subset of previous data. This setup simulates real-world constraints such as limited storage and privacy, requiring the model to maintain performance across diverse domains without catastrophic forgetting.

3.2 Benchmark Construction

Data Collection. We crawl full-text PDFs from the China National Knowledge Infrastructure (CNKI)[3], the largest Chinese academic repository. To capture broad domain diversity, we target 10

[3] https://www.cnki.net/

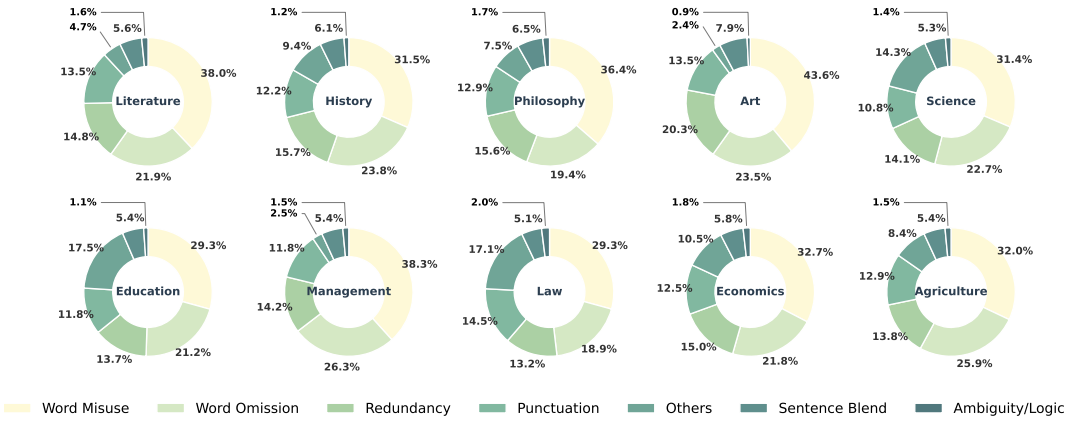

Figure 2: Error Type Statistics by First-Level Discipline.

first-level disciplines: *Law, Management, Education, Economics, Science, History, Agriculture, Literature, Art*, and *Philosophy*. The dataset is designed to be highly diverse and multi-disciplinary, ensuring comprehensive coverage of academic writing.

These first-level disciplines are further organized into 100 second-level disciplines to provide fine-grained coverage of academic writing. Beyond disciplinary coverage, we also examine how error patterns vary across domains. Figure 2 summarizes the composition of error types within each first-level discipline, illustrating both commonalities and discipline-specific differences in the distribution of error categories.

For each discipline, we randomly sample 1,000 sentences, yielding a balanced corpus of 10,000 instances. This one-to-one ratio eliminates domain-size bias and guarantees that subsequent continual-learning curricula are not dominated by any single field. To ensure a standardized evaluation, we provide a canonical split for each discipline: 700 training examples, 100 development examples, and 200 test examples.

Data Cleaning. Because CNKI only provides PDF files, a dedicated preprocessing pipeline is required. The overall cleaning procedure is executed by a trained annotation team and guarantees that every remaining sentence is grammatically self-contained and suitable for correction.

1. **PDF → JSON conversion**. We convert each PDF into a structured JSON file that preserves sentence boundaries, section tags, and positional metadata. This machine-readable format facilitates downstream filtering and reproducibility.

2. **Section filtering**. Only the *abstract* and *main body* are retained. Other sections like references, acknowledgements are discarded. These sections contain the bulk of scientific exposition and therefore the majority of grammar-related errors relevant to writing assistance.

3. **Sentence segmentation**. The retained text is split into sentence-level units using *LTP* (Che et al., 2021), enabling sentence-level GEC evaluation.

4. **Noise removal**. Inline citations, sub-titles, mathematical equations, tables, figures, and their captions are stripped. Eliminating non-linguistic tokens avoids misleading the error-detection models and prevents annotators from spending time on irrelevant content.

5. **Anonymisation**. All personal, institutional, and document identifiers are redacted to comply with privacy regulations and facilitate open release.

Data Annotation. Given the low error density in scholarly writing, fully manual annotation would be prohibitively expensive. We therefore adopt a human-in-the-loop strategy that combines automatic grammatical error detection, LLM pre-correction, human annotation, and expert validation.

1. **Automatic grammatical error detection**. 6 well-trained CGEC models (including GECToR (Omelianchuk et al., 2020) and fine-tuned Chinese-BART (Shao et al., 2021)) are first applied to the cleaned corpus. Only sentences flagged *consistently* by *all* detectors are kept.

Split	#Samples	Samples with Errors (%)	Avg Src Len.	Avg Ref Len.	Avg #Refs	Avg Edit Dist.
Train	7,000	6,456 (92.23%)	81.6	81.5	1.06	7.85
Dev	1,000	994 (99.40%)	63.4	63.7	1.94	9.61
Test	2,000	1,983 (99.15%)	64.0	63.9	1.95	9.60

Table 1: Basic Statistics of CL^{2}GEC.

This consensus voting filters out roughly 95% of grammatically correct lines, concentrating annotation effort on the 5% most error-prone candidates and dramatically cutting both LLM invocation and human labor.

2. **LLM pre-correction**. The shortlisted sentences are passed to GPT-4o (Hurst et al., 2024), which produces a candidate correction for each error span. These machine suggestions serve as weak references, giving annotators a standardising correction style across workers.

3. **Human annotation**. We recruit senior undergraduates or graduates with majors that *match* the corresponding discipline, ensuring domain awareness. After extensive training on pilot samples, annotators correct each sentence while consulting the detector output and GPT-4o suggestions. Every instance is independently revised by at least two annotators, which both improves recall and exposes stylistic alternatives.

4. **Expert validation**. Domain experts (including the paper authors) perform 100% manual review of the double-annotated data. They refine erroneous edits, reconcile conflicts, and may add supplementary references when multiple acceptable rewrites exist. The outcome is a high-precision, multi-reference gold annotation set.

This multi-stage pipeline maximises annotation quality *and* cost-effectiveness: automatic error detection minimises wasted effort, LLM pre-correction accelerates human editing, dual annotation guarantees inter-annotator agreement, and expert review delivers publication-grade reliability. Table 1 reports the basic statistics of CL^{2}GEC, including the proportion of grammatical-error instances, average sentence lengths, the number of references, and edit distance.

3.3 Evaluation Procedure and Metrics

We evaluate model performance in a continual learning setting using both standard GEC metrics and continual learning metrics.

3.3.1 Evaluation Protocol

Let $\{T_1, \ldots, T_N\}$ denote the sequence of $N = 10$ tasks, corresponding to our ten academic disciplines. In this continual learning setup, models are trained sequentially on discipline-specific training sets. To assess robustness to domain shift, we consider two curricula: (i) a semantically ordered sequence and (ii) a randomly shuffled sequence.

After learning each task T_i, the model is evaluated on all tasks from T_1 up to T_i. We record the following scores: (1) $Q_{i,i}$, the performance on the current task T_i immediately after training; (2) $Q_{i,j}$ ($j < i$), the performance on a past task T_j after training on T_i; and (3) $Q_{N,j}$, the final performance on each task T_j after completing the entire sequence of N tasks.

3.3.2 Standard GEC Metrics

We evaluate grammatical error correction (GEC) performance using the ChERRANT scorer (Zhang et al., 2022), which extends ERRANT to Chinese by performing character-level alignment and edit classification. For each evaluation result $Q_{i,j}$, ChERRANT computes Precision (P), Recall (R), and $F_{0.5}$ by classifying edits as true positives, false positives, or false negatives.

To summarize performance across tasks, we compute averages independently for each metric $M \in \{P, R, F_{0.5}\}$:

$$\overline{Q}^{(M)} = \frac{1}{N} \sum_{j=1}^{N} Q_{N,j}^{(M)}. \tag{4}$$

Here $Q_{N,j}^{(M)}$ denotes the final score on task T_j under metric M. Importantly, $F_{0.5}$ is averaged directly across tasks rather than recomputed from the averaged P and R.

3.3.3 Continual Learning Metrics

To capture retention and overall competence under sequential training, we adopt two standard continual learning metrics, computed separately for each $M \in \{P, R, F_{0.5}\}$.

Backward Transfer (BWT). BWT measures the average change in performance on past tasks from immediately after training to the end of the sequence. Negative values indicate forgetting, while non-negative values indicate retention or positive transfer:

$$\text{BWT}^{(M)} = \frac{1}{N-1} \sum_{j=1}^{N-1} \left(Q_{N,j}^{(M)} - Q_{j,j}^{(M)} \right). \quad (5)$$

Average Task Performance (AvgPerf). AvgPerf measures the model's overall GEC ability across all tasks after completing the full training sequence:

$$\text{AvgPerf}^{(M)} = \frac{1}{N} \sum_{j=1}^{N} Q_{N,j}^{(M)}. \quad (6)$$

4 Experiments

4.1 Experimental Settings

Backbones. All experiments use Qwen2.5-7B-Instruct (Qwen et al., 2025) and Llama3-8B-Instruct (Llama Team, 2024) as backbone models, chosen for their strong Chinese-language performance and robust instruction-following capabilities.

Task Sequences and Evaluation. To comprehensively assess the impact of task order on continual learning performance, we conducted experiments using two distinct training sequences:

1. **Randomized Order**: The 10 academic disciplines are presented in a randomly shuffled order. To ensure robustness and account for variance, this process is repeated across 5 different random permutations. We report the average results across these runs.

2. **Semantically Similar Order**: The tasks are arranged based on their semantic similarity. This sequence simulates a smoother domain transition and is used to investigate the effect of gradual domain shift on catastrophic forgetting.The definition and computation of similarity, as well as the resulting task order, are detailed in the Appendix.

4.2 Continual Learning Methods

We investigate the performance of various continual learning methods applied to the GEC task. Given that our task requires adapting a large language model to a series of distinct yet related domains, we focus on a strategy combining Parameter-Efficient Tuning (PET) with established continual learning algorithms. To this end, we benchmark the following four categories of adaptation strategies:

- **Sequential Finetuning (SeqFT)**: A naive baseline where the model is trained on each task in sequence without any specific mechanism to retain prior knowledge. This approach provides a lower bound for performance and highlights the problem of catastrophic forgetting.

- **Parameter-Efficient Tuning (LoRA)**: We apply Low-Rank Adaptation with rank 8. This serves as a lightweight adaptation approach.

- **Replay-based Methods**: To mitigate forgetting, we implement experience replay by retaining 2%, 5%, or 10% of training data from previous tasks.

- **Continual Learning Algorithms**: For our GEC task, we combine Parameter-Efficient Tuning (LoRA) with a set of representative continual learning algorithms to achieve superior results. We evaluate four such approaches: **EWC** (Elastic Weight Consolidation) (Kirkpatrick et al., 2017), which regularizes important parameters; **LwF** (Learning without Forgetting) (Li and Hoiem, 2016), which uses knowledge distillation; **GEM** (Gradient Episodic Memory) (Lopez-Paz and Ranzato, 2017), which constrains gradient updates; **OGD** (Orthogonal Gradient Descent) (Farajtabar et al., 2019), which minimizes task interference through orthogonal updates.

4.3 Non-Continual Baselines

To better contextualize the gains from continual learning, we additionally evaluate two settings that do not operate under the sequential CL constraint: (1) a raw one-shot baseline using a fixed, domain-agnostic GEC instruction without any task adaptation, and (2) a joint multi-domain LoRA baseline trained once on the mixture of all 10 disciplines under the same setting as our LoRA-based CL models. Raw prompting provides a low adaptation floor, while joint multi-domain training serves as an upper reference when all disciplines are simultaneously available. Detailed results are reported in Appendix.

5 Analysis

5.1 Overall Performance

Across both backbones and task orders, continual learning (CL) strategies clearly outperform sequential fine-tuning and generally surpass LoRA, while

Model	Strategy	GEC (Rnd)			GEC (Sem)			AvgPerf (Rnd)			AvgPerf (Sem)			BWT (Rnd)			BWT (Sem)		
		P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$
Qwen2.5 7B-Instruct	SeqFT	59.25	10.71	29.91	52.10	12.61	31.08	50.92	11.97	29.57	46.70	14.84	31.18	**8.13**	-1.20	-0.40	0.95	-0.13	0.63
	LoRA	65.42	13.00	35.54	64.18	13.03	34.83	62.80	11.33	32.02	61.70	12.92	33.94	4.54	1.61	4.01	1.77	**1.55**	**3.00**
	Replay	58.78	11.49	31.13	47.04	11.22	26.90	50.85	12.58	29.93	48.00	13.89	30.89	5.75	-1.73	-0.65	-4.31	-1.64	-4.17
	EWC	67.34	13.00	35.52	64.26	12.97	34.76	**64.88**	11.40	32.11	61.70	12.93	33.94	4.44	1.60	4.00	1.94	1.40	2.77
	GEM	67.33	13.10	**35.65**	**64.34**	13.00	34.82	64.86	11.42	32.17	61.77	12.93	33.96	4.41	**1.70**	**4.11**	**1.99**	1.45	2.84
	LwF	62.86	**13.22**	34.89	63.73	12.98	34.75	58.60	**12.84**	**33.36**	61.36	13.84	35.24	3.10	0.01	0.69	0.95	0.50	0.96
	OGD	**67.78**	12.30	34.44	62.77	**14.43**	**36.53**	63.15	13.38	34.97	**62.07**	**15.18**	**37.08**	4.85	-1.32	-0.81	-1.71	0.78	0.61
LLaMA3 8B-Instruct	SeqFT	45.40	9.10	24.14	45.60	9.98	26.01	35.15	10.73	22.52	36.56	11.03	24.00	6.11	-2.02	-1.32	4.03	-0.24	0.78
	LoRA	64.21	11.11	31.62	56.80	12.98	33.03	58.93	11.36	30.98	56.32	12.85	32.32	7.51	-0.22	0.64	-1.67	1.97	2.62
	Replay	37.31	10.07	23.70	34.32	9.56	22.17	30.00	10.39	20.79	27.90	10.77	20.46	7.41	-0.60	**2.50**	**7.73**	-0.52	2.64
	EWC	63.82	11.00	31.34	57.06	13.11	33.29	58.80	11.33	30.91	56.35	**12.91**	**32.38**	7.63	-0.22	0.67	-2.34	1.96	2.55
	GEM	**64.46**	11.12	31.71	57.94	13.16	33.53	**58.97**	11.35	30.99	56.28	12.82	32.26	**8.13**	-0.12	0.96	-0.67	2.19	3.16
	LwF	59.75	11.18	30.98	**59.79**	10.18	29.80	57.24	11.60	30.86	**58.99**	11.56	31.29	3.04	-0.43	0.06	0.09	0.87	1.53
	OGD	60.37	**12.89**	**33.54**	57.99	**13.42**	**33.89**	57.12	**12.92**	**32.74**	56.56	12.74	32.37	3.15	**-0.05**	0.49	-1.07	**2.30**	**2.92**

Table 2: Main Results of CL Strategies on CL^{2}GEC, Random (Rnd) vs. Semantic (Sem) Order.

Model	Buffer	GEC (Rnd)			GEC (Sem)			AvgPerf (Rnd)			AvgPerf (Sem)			BWT (Rnd)			BWT (Sem)		
		P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$	P	R	$F_{0.5}$
Qwen2.5 7B-Instruct	2 %	56.69	11.15	30.02	**50.56**	12.15	**30.12**	49.56	12.27	29.25	47.94	**14.44**	**31.27**	5.39	-1.26	**-0.61**	**-0.54**	-0.68	-0.73
	5 %	**58.78**	**11.49**	**31.13**	47.04	11.22	26.90	**50.85**	**12.58**	**29.93**	48.00	13.89	30.89	**5.75**	-1.73	-0.65	-4.31	-1.64	-4.17
	10 %	53.45	11.48	29.92	47.51	**12.18**	29.13	49.45	12.24	29.50	**48.56**	13.01	30.21	2.40	**-1.13**	-0.64	-2.97	**-0.67**	**-0.59**
LLaMA3 8B-Instruct	2 %	**45.24**	8.60	**23.71**	**43.47**	9.33	**24.62**	**33.02**	10.37	**21.68**	**34.71**	**11.00**	**23.50**	**13.36**	-2.06	1.32	**9.20**	-1.10	1.31
	5 %	37.31	**10.07**	23.70	34.32	**9.56**	22.17	30.00	**10.39**	20.79	27.90	10.77	20.46	7.41	**-0.60**	**2.50**	7.73	**-0.52**	**2.64**
	10 %	37.25	8.85	22.06	33.54	8.4	20.62	29.36	9.41	19.93	30.61	10.02	21.01	9.38	-1.03	2.11	4.26	-0.94	0.74

Table 3: Replay Results on CL^{2}GEC Benchmark.

Replay remains unreliable. This highlights the particular sensitivity of grammatical error correction (GEC) to catastrophic forgetting and the necessity of mechanisms that explicitly preserve previously acquired knowledge. Among the backbones, Qwen2.5-7B-Instruct consistently achieves higher scores than LLaMA3-8B-Instruct, suggesting that multilingual pretraining provides stronger inductive biases for this task; nevertheless, the relative ordering of methods remains stable across models, underscoring the robustness of the observed trends. LoRA serves as a strong baseline by constraining updates to a low-rank subspace, which alleviates the most severe forgetting, but its stability decreases under semantically ordered tasks, showing that parameter-efficient adaptation alone cannot fully prevent drift. CL methods systematically close this gap by directly addressing task interference, yielding more consistent performance across domains and curricula.

5.2 Continual Learning Strategies

While CL methods uniformly outperform the baselines, their strengths diverge systematically according to algorithmic design.

Projection-based methods (OGD) emphasize forward plasticity. By enforcing gradient orthogonality, OGD effectively incorporates new error patterns from diverse domains and achieves the strongest overall performance. This comes at the expense of weaker backward transfer, reflecting its bias toward adaptability rather than long-term retention.

Constraint-based methods (GEM, EWC) prioritize stability. GEM attains the strongest backward transfer, consistent with its objective of constraining updates to protect earlier tasks—a natural fit for GEC, where many grammatical structures recur across domains. EWC provides a more balanced trade-off, maintaining competitive final performance while offering robust retention.

Distillation-based methods (LwF) provide moderate but consistent gains. By distilling predictions from earlier tasks, LwF reduces forgetting

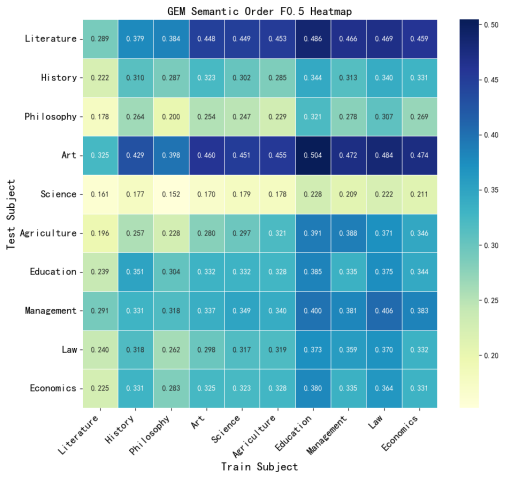

Figure 3: GEM Semantic Order ($F_{0.5}$) Heatmap.

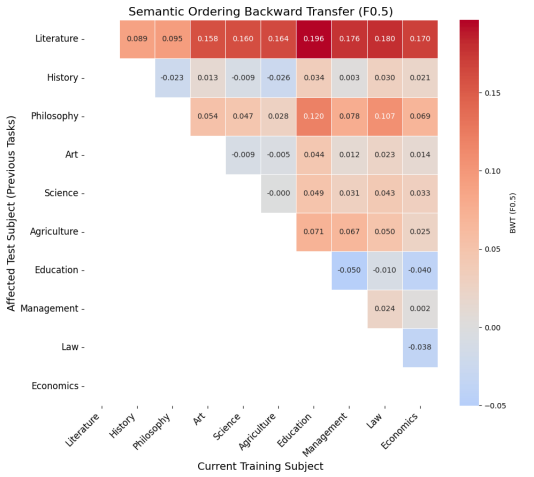

Figure 4: Backward Transfer ($F_{0.5}$) with Semantic Ordering.

without requiring memory storage. It consistently improves over LoRA, though generally trails OGD and GEM in overall performance.

Replay is less effective in this setting. Its limitation is not simply due to a small memory buffer. As shown in Table 3, replay exhibits a non-monotonic trend as the buffer size increases, indicating that its effectiveness is sensitive to buffer composition, buffer size, and task order. A plausible explanation is that larger replay buffers may introduce stronger cross-discipline interference and blur correction boundaries, which can hurt precision and thus $F_{0.5}$ even when replay coverage increases. Constraint- or projection-based methods therefore appear more effective and more stable than raw memory in this benchmark.

5.3 Replay Strategy

We further analyze the replay strategy by varying the buffer size (2%, 5%, 10%) on both Qwen2.5-7B-Instruct and LLaMA3-8B-Instruct (Table 3). The results reveal that replay does not scale monotonically with buffer size; instead, its effectiveness peaks at small to medium buffers and is strongly model-dependent.

5.4 Impact of Task Order

Task order systematically affects continual learning. Under semantic curricula, most strategies achieve higher recall and $F_{0.5}$ but lower precision. This trend is consistent in GEM, EWC, and Replay, implying that consecutive exposure to related tasks promotes broader generalization (improving coverage) at the cost of specificity (reducing precision), potentially due to blurrier correction boundaries within similar domains. In contrast, random orders increase topical diversity, which may sharpen task boundaries and thus improve precision, but typically weakens cross-task generalization.

Effects on backward transfer (BWT) are model-dependent. For Qwen2.5-7B-Instruct, semantic ordering often decreases BWT, suggesting that clustered similar tasks can exacerbate interference and hinder reuse of earlier knowledge. Conversely, LLaMA3-8B-Instruct frequently shows improved BWT under semantic order (e.g., with GEM and LoRA), indicating that redundancy across related tasks may help consolidate less robust representations. Overall, task-order effects are mediated by both model inductive biases and the CL mechanism.

5.5 Case Study: GEM under Different Task Orders

To better understand the dynamics of continual learning, we conduct a detailed case study of the GEM strategy using $F_{0.5}$ and BWT heatmaps (Figures 3 and 4). These visualizations provide a fine-grained view of how knowledge is preserved or overwritten as training progresses across tasks. For clarity, we present only the semantic-order results in the main text; the corresponding random-order heatmaps are included in the Appendix.

$F_{0.5}$ performance stability. The heatmap in Figure 3 tracks $F_{0.5}$ scores for each test subject (y-axis) after successive training tasks (x-axis, ordered from left to right). GEM maintains strong and relatively stable performance across the sequence, confirming

its effectiveness in mitigating catastrophic forgetting. Compared to random ordering (Appendix, Figure 5), semantic ordering yields smoother performance transitions, indicating that related tasks reinforce one another and support more predictable accumulation of knowledge.

Backward transfer dynamics. Figure 4 shows the BWT matrix, where the x-axis denotes the current training task and the y-axis denotes the previously learned test task. Semantic ordering promotes strong positive transfer among related domains. This illustrates how semantically structured curricula can leverage domain synergies to strengthen prior knowledge. At the same time, semantic ordering exhibits localized vulnerability: tasks less aligned with the curriculum suffer persistent negative BWT, indicating systematic forgetting. In contrast, random ordering (Appendix, Figure 6) yields more scattered and less severe negative BWT values, reducing the likelihood of any single task being consistently overwritten, though at the cost of weaker and less predictable positive transfer.

6 Conclusion

We introduced CL^{2}GEC, the first continual learning benchmark for Chinese grammatical error correction (CGEC). CL^{2}GEC simulates domain-incremental learning through a 10-discipline corpus of 10,000 human-annotated sentences, enabling sequential training and fine-grained evaluation of forgetting, adaptation, and transfer. We defined tailored evaluation protocols and benchmarked strong baselines using parameter-efficient tuning and four representative CL algorithms. Results show that regularization- and projection-based methods outperform sequential fine-tuning and replay, though performance varies with task order and model backbone. We hope CL^{2}GEC provides a foundation for future work on adaptive GEC and inspires the development of lifelong writing assistants capable of generalizing across academic domains.

7 Limitations

Reliance on a Specific Type of Continual Learning. This work primarily evaluates regularization-based continual learning methods like OGD and GEM. While these methods show strong performance, other approaches such as memory-augmented or meta-learning-based methods might offer complementary benefits. Future research could explore these alternatives for a more comprehensive understanding of continual learning in CGEC.

Generalization to Other Languages. Our experiments focus on Chinese GEC, and the methods proposed may not directly generalize to other languages with different grammatical structures. Further research is needed to assess the applicability of CL^{2}GEC to languages other than Chinese.

Evaluation on Limited Model Architectures. We evaluate two large language models, Qwen2.5-7B-Instruct and LLaMA3-8B-Instruct. However, other model architectures, including smaller models or domain-specific models, may yield different results. Expanding the evaluation to include a wider range of architectures would provide a broader perspective on the effectiveness of continual learning methods.

8 Ethics Statement

In this paper, we introduce the CL^{2}GEC benchmark, which is constructed from a custom-curated dataset. We have carefully detailed the collection, preprocessing, and annotation processes to ensure that no unethical behavior or infringement occurred during the dataset construction. To comply with ethical standards, we focus on data anonymization, desensitization, and the removal of any potentially harmful or biased content. The texts used in our dataset are sourced from publicly available academic materials, ensuring that the research tasks and directions proposed do not cause harm to society.

Acknowledgements

This research is supported by National Natural Science Foundation of China (Grant No.62276154); the Natural Science Foundation of Guangdong Province (Grant No.2024TQ08X729); Basic Research Fund of Shenzhen City (Grant No.JCYJ20240813112009013 and GJHZ20240218113603006); The Major Key Project of PCL for Experiments and Applications (Grant No.PCL2024A08).

References

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. N-LTP: An open-source neural language technology platform for Chinese. In *Proceedings of*

the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 42–49, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385.

Mehrdad Farajtabar, Navid Azizan, A. Mott, and Ang Li. 2019. Orthogonal gradient descent for continual learning. *Cornell University - arXiv,Cornell University - arXiv*.

Changhao Guan, Chao Huang, Hongliang Li, You Li, Ning Cheng, Zihe Liu, Yufeng Chen, Jinan Xu, and Jian Liu. 2025. Multi-stage LLM fine-tuning with a continual learning setting. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5484–5498, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinghan He, Haiyun Guo, Kuan Zhu, Zihan Zhao, Ming Tang, and Jinqiao Wang. 2024. SEEKR: Selective attention-guided knowledge retention for continual learning of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3254–3266, Miami, Florida, USA. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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*, page 3521–3526.

Zixiao Kong and Xianquan Wang. 2025. Scholargec: Enhancing controllability of large language model for chinese academic grammatical error correction. *AAAI Conference on Artificial Intelligence*.

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 Computing Surveys*, 57(8):1–36.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024a. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.

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

Yinghui Li, Shang Qin, Jingheng Ye, Haojing Huang, Yangning Li, Libo Qin, Xuming Hu, Wenhao Jiang, Hai-Tao Zheng, and Philip S Yu. 2024b. Rethinking the roles of large language models in chinese grammatical error correction. *arXiv preprint arXiv:2402.11420*.

Yinghui Li, Qingyu Zhou, Yangning Li, Zhongli Li, Ruiyang Liu, Rongyi Sun, Zizhen Wang, Chao Li, Yunbo Cao, and Hai-Tao Zheng. 2022. 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, Dublin, Ireland, May 22-27, 2022*, pages 3202–3213. Association for Computational Linguistics.

Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S Yu. 2024c. When llms meet cunning texts: A fallacy understanding benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:112433–112458.

Yutong Li, Lu Chen, Aiwei Liu, Kai Yu, and Lijie Wen. 2024d. Chatcite: Llm agent with human workflow guidance for comparative literature summary. *arXiv preprint arXiv:2403.02574*.

Zhizhong Li and Derek Hoiem. 2016. *Learning without Forgetting*, page 614–629.

AI @ Meta Llama Team. 2024. The llama 3 herd of models.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Neural Information Processing Systems,Neural Information Processing Systems*.

Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Yangning Li, 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, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 576–589. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not

rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the llmification of chi: Unpacking the impact of llms at chi through a systematic literature review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Lv Qingsong, Yangning Li, Zihua Lan, Zishan Xu, Jiwei Tang, Yinghui Li, Wenhao Jiang, Hai-Tao Zheng, and Philip S Yu. 2025. Raise: Reinforenced adaptive instruction selection for large language models. *arXiv preprint arXiv:2504.07282*.

Mengyang Qiu, Qingyu Gao, Linxuan Yang, Yang Gu, Tran Minh Nguyen, Zihao Huang, and Jungyeul Park. 2025. Chinese grammatical error correction: A survey. *arXiv preprint arXiv:2504.00977*.

Mujahid Ali Quidwai, Chunhui Li, and Parijat Dube. 2023. Beyond black box ai-generated plagiarism detection: From sentence to document level. *arXiv preprint arXiv:2306.08122*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. 2023. Progressive prompts: Continual learning for language models.

Chenze Shao and Yang Feng. 2022. Overcoming catastrophic forgetting beyond continual learning: Balanced training for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2023–2036, Dublin, Ireland. Association for Computational Linguistics.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. 2017. Incremental learning of object detectors without catastrophic forgetting. In *Proceedings of the IEEE international conference on computer vision*, pages 3400–3409.

Yue Wang, Zilong Zheng, Juntao Li, Zhihui Liu, Jinxiong Chang, Qishen Zhang, Zhongyi Liu, Guannan Zhang, and Min Zhang. 2024. Towards more realistic Chinese spell checking with new benchmark and specialized expert model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16570–16580, Torino, Italia. ELRA and ICCL.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*.

Liu Xiao, Li Ying, and Yu Zhengtao. 2024. Chinese grammatical error correction via large language model guided optimization training. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1366–1380, Taiyuan, China. Chinese Information Processing Society of China.

Peng Xing, Yinghui Li, Shirong Ma, Xinnian Liang, Haojing Huang, Yangning Li, Hai-Tao Zheng, Wenhao Jiang, and Ying Shen. 2024. Mitigating catastrophic forgetting in multi-domain chinese spelling correction by multi-stage knowledge transfer framework. *arXiv preprint arXiv:2402.11422*.

Lvxiaowei Xu, Jianwang Wu, Jiawei Peng, Jiayu Fu, and Ming Cai. 2022. FCGEC: Fine-grained corpus for Chinese grammatical error correction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1900–1918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, 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*, pages 6174–6189, Singapore. Association for Computational Linguistics.

Jingheng Ye, Shang Qin, Yinghui Li, Xuxin Cheng, Libo Qin, Hai-Tao Zheng, Ying Shen, Peng Xing, Zishan Xu, Guo Cheng, et al. 2025a. Excgec: A benchmark for edit-wise explainable chinese grammatical error correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25678–25686.

Jingheng Ye, Shang Qin, Yinghui Li, Hai-Tao Zheng, Shen Wang, and Qingsong Wen. 2025b. Corrections

meet explanations: A unified framework for explainable grammatical error correction. *arXiv preprint arXiv:2502.15261*.

Jingheng Ye, Zishan Xu, Yinghui Li, Xuxin Cheng, Linlin Song, Qingyu Zhou, Hai-Tao Zheng, Ying Shen, and Xin Su. 2024. Cleme2. 0: Towards more interpretable evaluation by disentangling edits for grammatical error correction. *arXiv preprint arXiv:2407.00934*.

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

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022. Mucgec: a multi-reference multi-source evaluation dataset for chinese grammatical error correction. *arXiv preprint arXiv:2204.10994*.

A Appendix

A.1 Data Breakdown and Secondary Disciplines

As described in the main text, the CL^{2}GEC benchmark is composed of 10 primary disciplines, each containing a number of secondary disciplines. Since the total number of secondary disciplines is large, we focus on the most prominent ones for visualization. Figures 9 and 10 show the distribution of the top 10 most frequent secondary disciplines within each primary discipline. To visualize this distribution, a stacked bar chart is provided, illustrating the count and proportion of these top 10 secondary disciplines within each primary discipline.

A.1.1 Error Type Statistics

Table 4 summarizes the distribution of error types across the 10 first-level disciplines. Overall, *Word Misuse* is the dominant category in every discipline (29.3%–43.6%), followed by *Word Omission* (18.9%–26.3%). Mid-frequency errors are mainly *Redundancy* (13.2%–20.3%) and *Punctuation Errors* (10.8%–14.5%), whereas discourse-heavy issues such as *Sentence Blend* (5.1%–7.9%) and *Ambiguity/Logic* (0.9%–2.0%) are comparatively rare.

Beyond this global pattern, the error-type mixtures are especially similar among disciplines. For example, the *Science–Agriculture* pair shows closely matched proportions for major categories (*Word Misuse*: 31.4% vs. 32.0%; *Word Omission*: 22.7% vs. 25.9%; *Punctuation Errors*: 10.8% vs. 12.9%). A similar consistency is observed within the *Education–Management–Law–Economics* cluster: *Word Omission* stays within 18.9%–26.3%, and *Punctuation Errors* remain in a narrow band (11.8%–14.5%), alongside consistently high *Word Misuse* (29.3%–38.3%). Likewise, the humanities-oriented *Literature–History–Philosophy–Art* group shares a common profile of high *Word Misuse* (31.5%–43.6%) and substantial *Redundancy* (14.8%–20.3%), with relatively stable *Punctuation Errors* (12.2%–13.5%).

A.1.2 Randomized Order

The 10 academic disciplines are presented in a randomly shuffled order. To ensure robustness and account for variance, this process is repeated across 5 different random permutations. We report the average results across these runs.

A.1.3 Semantically Similar Order

The tasks are arranged based on the semantic similarity between the disciplines. The resulting order is designed to reflect a smoother transition between related domains, thus investigating how gradual shifts between similar tasks affect the model's ability to retain knowledge. The semantic similarity between disciplines is computed using sentence embeddings generated by the `SentenceTransformer` model. The average similarity matrix between the disciplines is presented in Table 5.

Semantic Similarity Computation We compute the semantic similarity between disciplines by first encoding each discipline's sentences into dense vector representations (embeddings) using the `SentenceTransformer` model and then averaged to produce a single vector for each discipline. We compute the cosine similarity between all pairs of discipline embeddings to measure how similar they are to each other in a semantic space.

The formula for cosine similarity between two vectors $\mathbf{v_1}$ and $\mathbf{v_2}$ is:

$$\text{Cosine Similarity} = \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{\|\mathbf{v_1}\|\|\mathbf{v_2}\|}$$

This cosine similarity value ranges from -1 (completely dissimilar) to 1 (identical).

Conclusion of Task Ordering Table 5 provides pairwise semantic similarities. To derive a reproducible semantic curriculum (an absolute task order), we use a deterministic two-step procedure: (i) grouping and (ii) ordering.

Discipline	Total Errors	Word Omission	Word Misuse	Redundancy	Punctuation Errors	Sentence Blend	Ambiguity/Logic	Others
Education	3,950	836 (21.2%)	1,159 (29.3%)	541 (13.7%)	468 (11.8%)	213 (5.4%)	43 (1.1%)	690 (17.5%)
Management	3,215	845 (26.3%)	1,232 (38.3%)	457 (14.2%)	379 (11.8%)	172 (5.4%)	48 (1.5%)	82 (2.5%)
History	2,890	688 (23.8%)	911 (31.5%)	454 (15.7%)	352 (12.2%)	177 (6.1%)	36 (1.2%)	272 (9.4%)
Law	3,420	645 (18.9%)	1,002 (29.3%)	452 (13.2%)	496 (14.5%)	173 (5.1%)	68 (2.0%)	584 (17.1%)
Science	4,105	930 (22.7%)	1,288 (31.4%)	579 (14.1%)	442 (10.8%)	218 (5.3%)	59 (1.4%)	589 (14.3%)
Philosophy	3,215	624 (19.4%)	1,171 (36.4%)	503 (15.6%)	414 (12.9%)	208 (6.5%)	56 (1.7%)	239 (7.5%)
Economics	3,450	751 (21.8%)	1,127 (32.7%)	517 (15.0%)	431 (12.5%)	201 (5.8%)	61 (1.8%)	362 (10.5%)
Agriculture	2,980	773 (25.9%)	953 (32.0%)	412 (13.8%)	386 (12.9%)	161 (5.4%)	44 (1.5%)	251 (8.4%)
Literature	2,715	595 (21.9%)	1,031 (38.0%)	401 (14.8%)	367 (13.5%)	151 (5.6%)	42 (1.6%)	128 (4.7%)
Art	2,145	503 (23.5%)	935 (43.6%)	436 (20.3%)	289 (13.5%)	170 (7.9%)	19 (0.9%)	52 (2.4%)

Table 4: Error Type Statistics.

Discipline	Literature	History	Philosophy	Education	Law	Science	Agriculture	Economics	Management	Art
Literature	1.0000	0.2115	0.2130	0.1506	0.1418	0.0493	0.0334	0.1185	0.1406	0.1923
History	0.2115	1.0000	0.2345	0.2252	0.2059	0.0872	0.0846	0.1733	0.2139	0.1769
Philosophy	0.2130	0.2345	1.0000	0.2024	0.2388	0.0923	0.0792	0.2173	0.1999	0.1553
Education	0.1506	0.2252	0.2024	1.0000	0.2324	0.1230	0.1340	0.1869	0.2988	0.1224
Law	0.1418	0.2059	0.2388	0.2324	1.0000	0.1036	0.1019	0.2111	0.2266	0.1223
Science	0.0493	0.0872	0.0923	0.1230	0.1036	1.0000	0.2022	0.1449	0.1523	0.0758
Agriculture	0.0334	0.0846	0.0792	0.1340	0.1019	0.2022	1.0000	0.1548	0.1508	0.0338
Economics	0.1185	0.1733	0.2173	0.1869	0.2111	0.1449	0.1548	1.0000	0.2093	0.0915
Management	0.1406	0.2139	0.1999	0.2988	0.2266	0.1523	0.1508	0.2093	1.0000	0.1286
Art	0.1923	0.1769	0.1553	0.1224	0.1223	0.0758	0.0338	0.0915	0.1286	1.0000

Table 5: Semantic Similarity Matrix between Disciplines

Step 1: Task Grouping We form three curriculum stages with group sizes (4,4,2), corresponding to a coarse-grained three-stage progression. Among all partitions of the 10 disciplines that satisfy these sizes, we select the partition that maximizes the sum of intra-group similarities (excluding diagonal entries). This yields the following groups:

- **G1:** Literature, History, Philosophy, Art
- **G2:** Education, Management, Law, Economics
- **G3:** Science, Agriculture

These groups are also supported by their higher intra-group similarities compared with inter-group similarities (e.g., G1–G3 is substantially lower than G1–G2 in Table 5).

Step 2: Absolute Ordering Within and Across Groups We order the three groups as G1 $\rightarrow$ G2 $\rightarrow$ G3, since G2 is semantically closer to both G1 and G3 than G1 is to G3 (Table 5), yielding a smoother transition. Within this fixed group order, we choose the within-group permutations that maximize the total adjacent similarity over the entire curriculum:

$$\max_{\pi} \sum_{t=1}^{9} \mathrm{sim}(\pi_t, \pi_{t+1}),$$

subject to π respecting the group membership. Because each group is small, we enumerate permutations within each group and pick the best one deterministically (ties are broken alphabetically). The resulting absolute semantic curriculum is:

$$\pi = (\text{Art}, \text{Literature}, \text{History}, \text{Philosophy}, \text{Law}, \text{Education}, \text{Management}, \text{Economics}, \text{Agriculture}, \text{Science}).$$

A.2 Implementation Details.

All experiments are conducted under a unified training setup. For the randomized curriculum, we report the average results over five random permutations, while replay-based methods use replay ratios of 0.02, 0.05, and 0.10. Detailed hyper-parameter configurations used in our experiments, including the training setup, replay settings, and decoding parameters, are summarized in Table 6.

A.2.1 Prompt for GPT-4o Pre-correction.

During dataset construction, GPT-4o is used to generate a candidate correction before human annotation. For reproducibility, we report the full prompt template used in our experiments in Figure 7. The upper part shows the original English prompt used in the experiments, and the lower part provides a Chinese translation for readability.

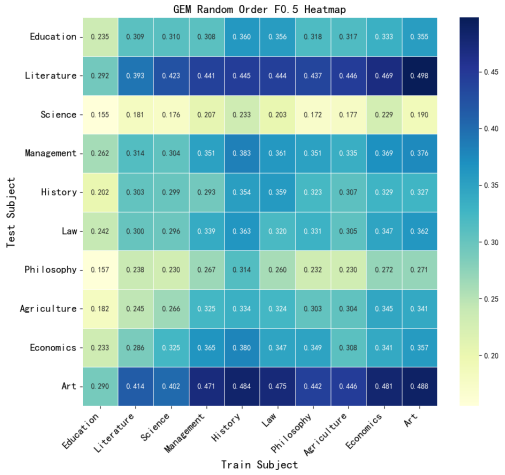

Figure 5: GEM Random Order ($F_{0.5}$) Heatmap.

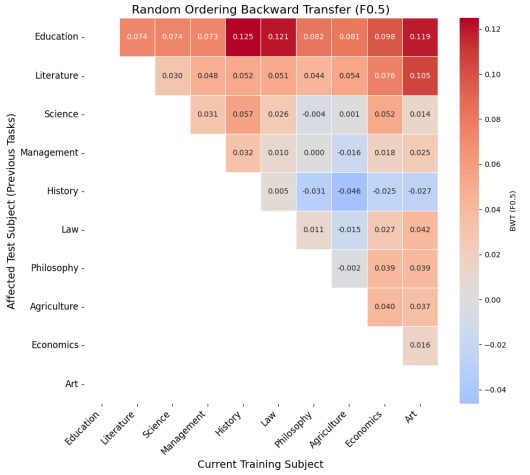

Figure 6: Backward Transfer ($F_{0.5}$) with Random Ordering.

Configuration	Value
Training	
Devices	8 Tesla A100 GPU (80GB)
Epochs	3
Training batch size per GPU	4
Evaluation batch size per GPU	16
Gradient accumulation steps	8
Learning rate	5×10^{-5}
Learning rate schedule	cosine decay
Warmup steps	20
Max prompt length	1024
Max answer length	512
Preprocessing workers number	16
Seed	1234
Zero stage	2
Replay	
Replay ratio	0.02 / 0.05 / 0.10
Inference	
Beam size	5
Top-p	0.8
Max new tokens	2048
Temperature	0.7

Table 6: Main hyper-parameter configuration used in our experiments.

A.2.2 Raw One-shot Baseline Template.

For the raw one-shot baseline, we use a fixed instruction template with one in-context example and no task-specific adaptation or finetuning. For reproducibility, Figure 8 presents the prompt template used in this setting. The upper part shows the original Chinese prompt used in the experiments, and the lower part provides an English translation for readability.

A.2.3 Non-Continual Baselines.

We evaluate two non-continual settings: raw one-shot inference and joint multi-domain LoRA finetuning. Table 7 summarizes their overall performance on CL^{2}GEC.

A.3 The Detail of Replay Strategy

Qwen2.5-7B-Instruct achieves its best performance at a 5% buffer under random order, but both GEC and AvgPerf decline when the buffer is enlarged to 10%. Moreover, backward transfer (BWT) remains consistently negative, with particularly sharp degradation under semantic order, suggesting that replay alone is insufficient to mitigate forgetting for Qwen2.5-7B-Instruct. By contrast, LLaMA3-8B-Instruct benefits more from replay, with its strongest GEC and AvgPerf observed at 2% buffer, and BWT peaking around 5% before dropping again at 10%. Overall, replay does not improve monotonically with larger buffers; instead, its effectiveness is strongly model-dependent and appears sensitive to buffer composition, buffer size, and task order. A plausible explanation is that larger replay buffers may introduce stronger cross-discipline interference and blur correction boundaries, which can hurt precision and thus $F_{0.5}$ even when replay coverage increases.

A.4 Case Study

We provide a case study of the CL^{2}GEC benchmark across 10 disciplines. The results are visualized in Figure 11– 15, showing representative examples across different academic domains.

Baseline setting	Backbone	P	R	$F_{0.5}$
Raw inference (one-shot)	LLaMA3-8B-Instruct	25.97	12.59	21.42
Raw inference (one-shot)	Qwen2.5-7B-Instruct	36.72	17.63	30.18
Joint multi-domain finetune (LoRA)	LLaMA3-8B-Instruct	53.60	14.73	35.08
Joint multi-domain finetune (LoRA)	Qwen2.5-7B-Instruct	56.31	16.23	37.69

Table 7: Raw and Joint Baselines on CL^{2}GEC.

English Prompt

You are a Chinese grammatical error correction tool. Your task is to correct the grammatical errors in an academic paper. Make the smallest possible change in order to make the content grammatically correct. Change as few words as possible. Do not rephrase parts of the content that are already grammatical. Do not change the meaning of the content by adding or removing information. If the content is already grammatically correct, you must output the original content without changing anything. You should output directly the target sentence and do not add extra explanations.

Here is the formal input. We include the context of the given content to help you understand the formal content if available. You should focus on the correction of the content rather than its context.

Category: `{category}`
Discipline: `{discipline}`
Title: `{title}`
Previous context: `{prev_context}`
Content: `{content}`
Next context: `{next_context}`
Output content:

中文提示词

你是一个中文语法纠错工具。你的任务是纠正学术论文中的语法错误。请以尽可能小的改动使文本在语法上正确，尽量少改动原句内容。不要对已经正确的部分进行改写，不要通过增删信息改变原句含义。如果文本本身已经没有语法错误，你必须原样输出输入文本，不做任何修改。请直接输出纠正后的句子，不要添加任何额外解释。

下面是正式输入。为了帮助你理解当前文本的正式语境，我们可能会提供上下文信息；但你应当聚焦于对当前文本本身的纠正，而不是对上下文进行修改。

学科门类：`{category}`
一级学科：`{discipline}`
论文标题：`{title}`
上文：`{prev_context}`
输入文本：`{content}`
下文：`{next_context}`
输出文本：

Figure 7: Prompt template for GPT-4o pre-correction.

中文提示词

你是一个中文语法错误纠正工具。你的任务是纠正输入句子中的语法错误，并尽可能少做修改以使句子语法正确。不要改写句子中已经语法正确的部分。不要通过添加或删除信息来改变句子的意思。如果句子已经语法正确，你应该直接输出原句，不要做任何改动，也不要做额外解释。

示例输入： 为了上述四个问题，本研究分别从语法纠错数据资源、数据增强、可解释性、评估度量四个反面提出解决方案。

示例输出： 针对上述四个问题，本研究分别从语法纠错数据资源、数据增强、可解释性和评估度量四个方面提出相应的解决方案。

正式输入：`{待纠正句子}`
正式输出：

English Prompt

You are a Chinese grammatical error correction tool. Your task is to correct the grammatical errors in the input sentence while making as few changes as possible so that the sentence becomes grammatically correct. Do not rewrite parts of the sentence that are already grammatical. Do not change the meaning of the sentence by adding or removing information. If the sentence is already grammatically correct, you should output the original sentence directly, without making any changes or adding any extra explanation.

Example Input: 为了上述四个问题，本研究分别从语法纠错数据资源、数据增强、可解释性、评估度量四个反面提出解决方案。

Example Output: 针对上述四个问题，本研究分别从语法纠错数据资源、数据增强、可解释性和评估度量四个方面提出相应的解决方案。

Input: `{sentence}`
Output:

Figure 8: Instruction template for the raw one-shot baseline.

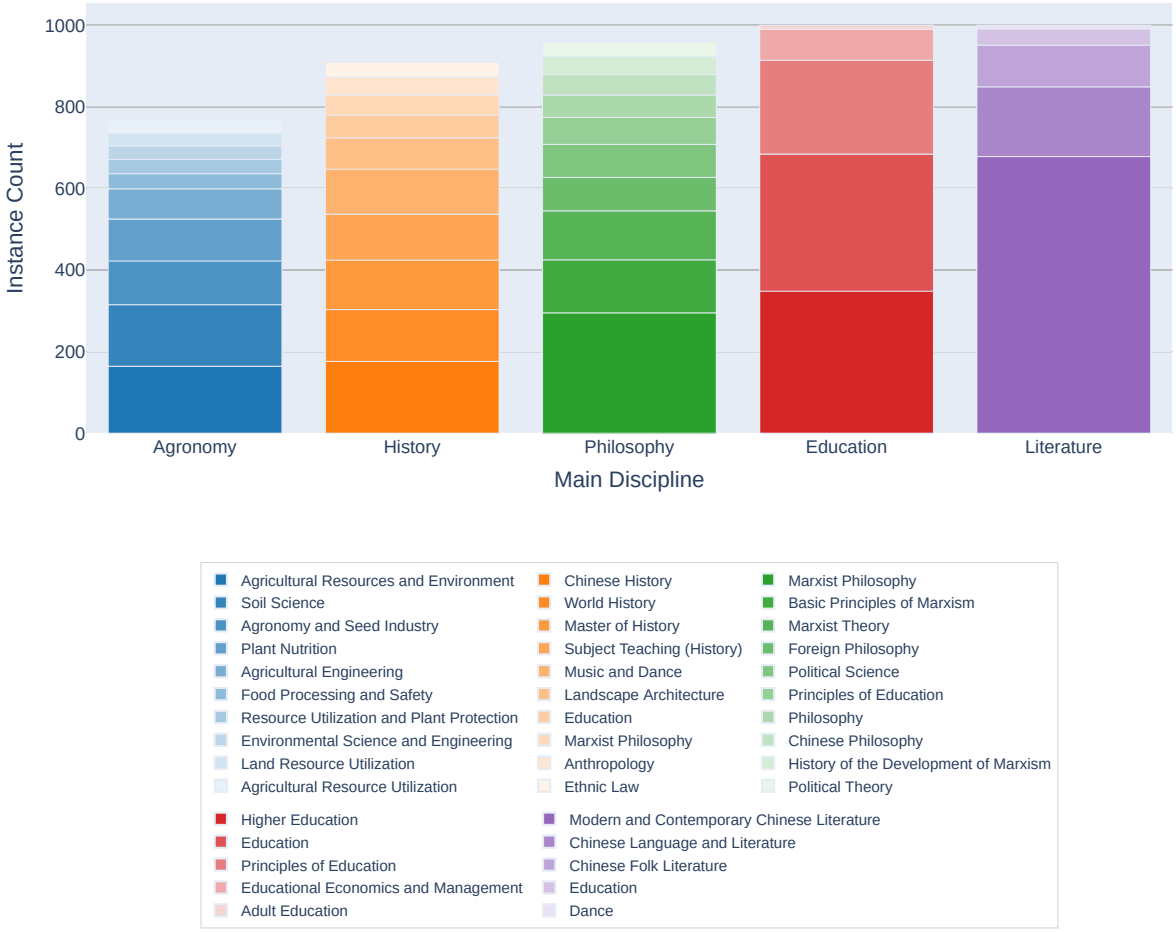

Figure 9: Discipline-wise Distribution of Top 10 Sub-disciplines (Part 1)

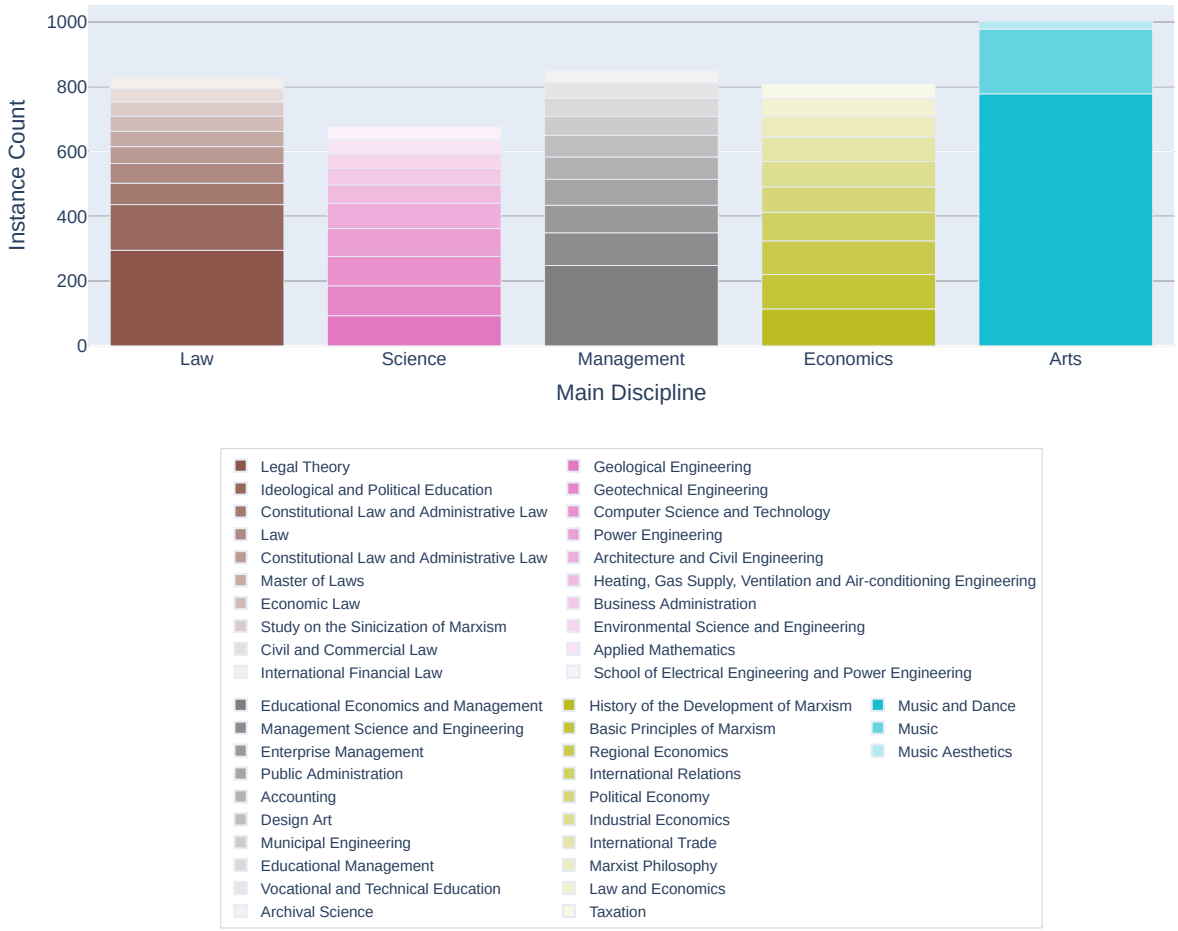

Figure 10: Discipline-wise Distribution of Top 10 Sub-disciplines (Part 2)

1. 理学

{

"input": "当气体中所含溶质存有浓度差，这类浓度的不平衡态将会引发物质转移，研究者把这类现象称之为扩散。",

"output": "当气体中所含溶质存有浓度差，这类浓度的不平衡态将会引发物质转移，研究者把这类现象称之为扩散。"

},
{

"input": "近年来，BiOI 作为较为新鲜的光催化材料，因其独特的层状结构和新颖的光学性能而被广泛研究。",

"output": "近年来，作为一种较为新鲜的光催化材料，BiOI 因其独特的层状结构以及新颖的光学性质受到了广泛的研究。"

}

2. 法学

{

"input": "在研究方法上，实证研究方法的类型包括社会调查方法、历史研究方法、比较研究方法、逻辑分析方法、语义分析方法。",

"output": "在研究方法上，实证研究方法包括社会调查方法、历史研究方法、比较研究方法、逻辑分析方法和语义分析方法。"

},
{

"input": "社科法学除了需要在自身的方法论研究、研究内容上做出反思与改进外，社科法学要想在学术界作为一个“学派”走得更远，就必须在组织上形成自己的学术共同体。",

"output": "社科法学除了需要在自身的方法论研究、内容上做出反思与改进外，社科法学要想在学术界作为一个“学派”走得更远，就必须在组织上形成自己的学术共同体。"

}

Figure 11: Examples of CL^2GEC.

3. 管理学

{

"input": "因此，为进一步提升学位论文的质量和水平，在学位论文选题中，既要考虑研究热点，也要考虑教育经济与管理学科范围内的本质问题、核心问题，而不能一味地去追求热点，只有这样研究成果才会被更多的人去关注，认可。",

"output": "所以，为了进一步提高学位论文的质量和水平，在选择学位论文题目时，既要关注研究热点，又要思考教育经济与管理学科范围内的本质问题、核心问题，不能盲目追求热点。这样，研究成果才会得到更多人的关注和认可。"

},
{

"input": "本文研究的意义在于探究式学习法这一较新的学习方式，可以使学生在探究和学习的过程中提高思考能力、学习能力、创新能力、协作能力，并对老师和学生之前的关系有着促进作用。",

"output": "本文研究的意义在于，探究式学习法作为一种较新的学习方式，能够让学生在探究和学习的过程中，提升思考能力、学习能力、创新能力和协作能力，并且对师生之间的关系起到促进作用。"

}

4. 教育学

{

"input": "高等教育学硕士研究生这样踏踏实实、勤勤恳恳的做学术研究、撰写论文，研究生三年实践扎扎实实学习积累，毕业前夕总能得到学术方面的“累累硕果”。 ",

"output": "高等教育学硕士研究生这样求真务实、兢兢业业地做学术研究、撰写论文，三年时间踏踏实实学习积累，毕业前夕就会得到学术方面的“累累硕果”。"

},
{

"input": "因为大学生是极具个性化的群体，因此大学生的发展表现出多样化，并极具个性差异，大学对于学生的管理要遵循学生的成长和发展规律。",

"output": "大学生是极具个性化的群体，表现出多样化并极具个性差异的特点，因此大学对于学生的管理要遵循学生的成长和发展规律。"

}

Figure 12: Examples of CL^2GEC.

5. 经济学

{

"input": "总揽历史长河，经济增长理论的发展共经历了三个不同的阶段，分别是古典增长理论，新古典增长理论和新增长理论。",

"output": "纵览历史长河，经济增长理论的发展共经历了三个不同的阶段，分别是古典增长理论、新古典增长理论和新增长理论。"

},
{

"input": "增加的资本要雇佣更多的劳动，如果此时劳动者不能从人口更多的国家转移过来，这种资本增加的趋势就会大大的提高劳动的市场价格。",

"output": "增加的资本要雇佣更多的劳动，如果此时劳动者不能从人口更多的国家转移过来，这种资本增加的趋势就会大大的提高劳动的市场价格。"

}

6. 历史学

{

"input": "学界大体认为，二战前为传统的'汉学研究'，二战后则为现代的'中国学研究'，两者都致力于中国语言、历史和文化的研究。",

"output": "学界大体认为，二战前为传统的"汉学研究"，二战后则为现代的"中国学研究"，两者都致力于中国语言、历史和文化的研究。"

},
{

"input": "二是对历史研学旅行课程反思的探究。当学生历经了一次完整的研学课程后，对于历史研学课程有切身体会，是否学到了、学到了什么等可以作为探究的范例，可以在回顾知识的同时加深对知识的理解，从而在以后的课程学习中，达到"顿悟"的境界；三是对旁支学科的探究，达到跨学科学习。",

"output": "二是对历史研学旅行课程反思的探究。当学生经历了一次完整的研学课程后，对于历史的研学课程有了切身体验。学生是否学到、学到了什么等可以作为探究的范例。可以在回顾知识的同时加深对知识的理解，从而在以后的课程学习中，达到"顿悟"的境界。三是对旁支学科的探究，达到跨学科学习。"

}

Figure 13: Examples of CL^{2}GEC.

7. 农学

{

"input": "由于土壤养分数据获取成本较高，需要数据量大，更新困难，应进一步探讨实时的养分数据更新方法，降低数据更新周期，提高其实时性。",

"output": "由于土壤养分数据获取成本较高、需要数据量大且更新困难，应进一步探讨实时的养分数据更新方法，以缩短数据更新周期，并提高其实时性。"

},
{

"input": "在花生上以及小麦上的研究也表现出类似的结果，施用硫肥可以改善作物生育性状，增加叶绿素含量，单株结果枝数、单株饱果数、百粒质量均有所增加，增产效果明显。",

"output": "在花生以及小麦上的研究也表现出类似的结果，施用硫肥可以改善作物生育性状，增加叶绿素含量，单株结果枝数、单株饱果数、百粒质量均有所增加，增产效果明显。"

}

8. 文学

{

“input": "画面中短衣帮们互相谈笑风生，穿长衫的则背着手傲气的走进酒店里，这一派景象十分鲜活灵动，富有生活气息。",

"output": "画面中短衣帮们互相谈笑风生，穿长衫的则背着手傲气地走进酒店里，这一派景象十分鲜活灵动，富有生活气息。"

},
{

"input": "有时他会陷入自我怀疑之中，哀叹“凡是一切顶小的顶平凡的生活事业，也不全是为我这样人而有的。我有的也许正是为人不屑要的。",

"output": "有时他会陷入自我怀疑之中，哀叹“凡是一切小而平凡的生活事业，也不全是为我这样人所拥有的。我有的也许是别人不屑要的。""

}

Figure 14: Examples of CL^{2}GEC.

9. 艺术学

{

"input": "单簧管继续演奏鸭子的主题，不同的是大提琴分奏变为三连音节奏型，在情绪上弥补了单簧管和大管的慢节奏，使其紧张性增加，作曲家想在此处在描写出鸭子跳入水中后动作变得灵活起来。",

"output": "单簧管继续演奏鸭子的主题，而大提琴的分奏转变为三连音节奏型。这种变化在情绪上弥补了单簧管和大管的慢节奏，增加了音乐的紧张感。作曲家意图通过这一改变来描绘鸭子跳入水中后动作变得更加灵活的场景。"

},
{

"input": "尖锐的音响效果与不和协因素的出现似乎在提醒和暗示战争带给人们的痛苦仍未散去。",

"output": "尖锐的音响效果与不和谐因素的出现似乎在提醒和暗示战争带给人们的痛苦仍未散去。"

}

10. 哲学

{

"input": "唯心主义将世界的本原归结为意识，这实质上模糊了世界本来的模样，那么既然唯心主义并没能使人正确的认识世界，那么唯心主义在历史长河中能够长久存在的真正原因需要揭露。",

"output": "唯心主义将世界的本原归结为意识，这实质上模糊了世界本来的模样，既然唯心主义并没能使人正确地认识世界，那么唯心主义在历史长河中能够长久存在的真正原因仍需要探索。"

},
{

"input": "既然意识之主体性在自身内部建构起来的对象世界终将归于抽象，意识也因此成为“非对象性的存在物”，即意识等于无。",

"output": "由于意识的主体性在其内部构建的对象世界终将归于抽象，因此意识也被认为是“非对象性的存在物”，也就是说，意识等于无。"

}

Figure 15: Examples of CL^{2}GEC.