LLM-SLM Collaborative Framework of Idiomatic Expression Generation

Hui Gao, Changhao Song, Peng Zhang*, Jing Zhang, Chang Yang, Liuxian Ge
School of Computer Science and Technology, Tianjin University, Tianjin, China
{hui_gao, pzhang}@tju.edu.cn

Abstract

Idiomatic Expression Generation, which aims to produce idiomatic text from plain text, is a valuable yet challenging NLP task. However, existing methods suffer from the scarcity of parallel data and dependence on high-quality manual annotations. To address this, we propose an iterative LLM-SLM (Large Language Model-Small Language Model) collaborative framework—Auto-IDEA, that replaces human supervision for idiomatic expression data generation. In this self-improving cycle, the LLM constructs parallel corpora (idiomatic and plain text) via bidirectional semantic reconstruction, automatically generating "Locate-Then-Polish" (LTP) annotations; the SLM filters low-quality corpora while continuously enhancing its verification ability through incremental learning. We instantiate Auto-IDEA for Chinese Idiom Polishing (CIP), constructing CIP-200K, a large-scale dataset of 206K parallel sentences with LTP annotations. The Qwen3-8B fine-tuned on CIP-200K achieves a 25.2% absolute Idiom Polishing Accuracy (IPA) improvement over a supervised fine-tuning (SFT) baseline, outperforming DeepSeek-R1 by 6.2%. Extensive experiments (e.g., Chinese idiom cloze tests and English idiom generation tasks) and human evaluations verify the generalization and effectiveness of Auto-IDEA, demonstrating a new pathway for high-quality, annotation-free data generation through LLM-SLM collaboration.

1 Introduction

Idiomatic expressions (e.g., "break the ice") are cornerstones of fluent communication. The task of **Idiomatic Expression Generation**—generating text containing idioms—is valuable for writing enhancement, language learning, and culturally-aware NLP (ERDAĞI, 2024; Zahara and Ekawati, 2025). However, it demands deep cultural understanding, which makes acquiring parallel training data difficult and evaluating quality challenging. Consequently, existing approaches often rely on manual annotation or are limited to selection-based tasks rather than open-ended generation (Ri et al., 2023; Gárate Rodas and Palacios Alvarado, 2025).

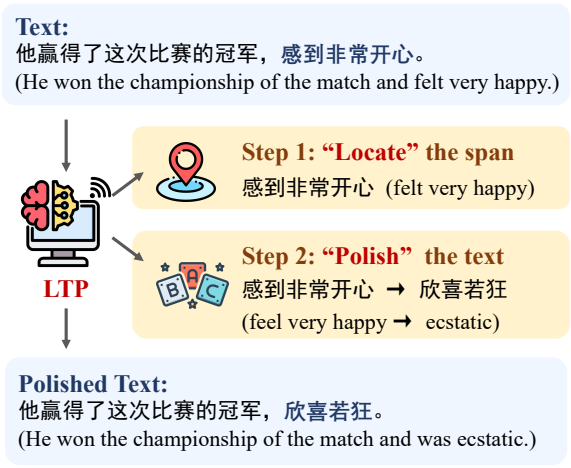

Figure 1: An example of the Chinese idiom polishing task, which needs to follow the reasoning process of "Locate-Then-Polish (LTP)" to ensure that the polishing task is precise and efficient.

Considering the cultural differences among various languages, we focus on Chinese Idiom Polishing (CIP) as a rigorous and representative task. Chinese idioms are conventionalized four-character expressions rich in cultural allusions, demanding exceptional semantic and cultural precision for proper use. We formulate CIP as a "Locate-Then-Polish" (LTP) process (as shown in Figure 1), which requires the model to first locate the span within a sentence that can be polished (e.g., "感到非常开心, felt very happy"), then select a contextually and culturally apt idiom (e.g., "欣喜若狂, ecstatic"), and seamlessly rewrite the text. While this explicit LTP formulation effectively decomposes the complex generation task, it creates a dependency on

*Corresponding author

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12125–12145

large-scale parallel data with LTP annotations. The core problem thus crystallizes: how to automatically generate such annotated data at scale to train models capable of this sophisticated reasoning?

To overcome this bottleneck, we propose **Auto-IDEA**, a novel unsupervised framework that generates high-quality idiomatic expression LTP annotations through iterative collaboration between a Large Language Model (LLM) and a Small Language Model (SLM). Auto-IDEA establishes a self-improving cycle: an LLM acts as the generator and annotator, producing diverse candidate sentences and, crucially, automatically deriving the corresponding LTP chain (i.e., the location and the idiom substitution) through bidirectional semantic reconstruction. The SLM serves as an adaptive verifier, filtering noisy candidates and undergoing incremental learning on the curated data. This process progressively refines the SLM's verification capability, closing the loop. Their collaboration eliminates the dependency on manual annotation while autonomously producing both the training data and the precise supervisory signals (LTP chains) needed to teach models the underlying reasoning.

Leveraging Auto-IDEA, we construct CIP-200K, a large-scale dataset for Chinese Idiom Polishing comprising 206,342 high-quality parallel sentences, each annotated with its LTP chain. Experiments on three mainstream LLMs (Qwen3, Llama3, GLM4) demonstrate that LTP fine-tuning significantly enhances their performance on CIP tasks. Notably, Qwen3-8B achieves 64.2% accuracy, surpassing DeepSeek-R1's performance. Furthermore, more extensive task experiments and comprehensive manual evaluations have demonstrated the generalization and robustness of this work. The main contributions of this work are threefold:

- **Novel Framework**: We propose Auto-IDEA, an iterative LLM-SLM collaborative framework that generates high-quality parallel data with **LTP annotations** autonomously, eliminating the need for manual supervision in idiomatic expression generation.
- **Benchmark Dataset**: We release CIP-200K, the first large-scale LTP-annotated dataset for Chinese Idiom Polishing, which provides a valuable benchmark for future research.
- **Performance improvement**: Extensive experiments demonstrate that models fine-tuned on CIP-200K achieve new state-of-the-art performance, surpassing the DeepSeek-R1 and proving the critical value of LTP annotation for idiomatic expression generation tasks.

2 Related Work

2.1 Idiomatic Expression Generation

Research on idiomatic expressions has long focused on comprehension tasks, such as cloze-style tests in the Chinese Idiom Dataset (ChID) (Zheng et al., 2019). While progress has been made via knowledge integration (Long et al., 2020; Wang et al., 2021), representation learning (Tan and Jiang, 2021; Sha et al., 2023), and contrastive learning (Wu et al., 2024), the more challenging idiomatic text generation task remains underexplored. Prior work related to generation includes idiom-aware machine translation (Shao et al., 2018; Li et al., 2024), paraphrase generation (Qiang et al., 2023) and other tasks (Wang et al., 2025; Wong et al., 2010; Pintado and Fajardo, 2021), all of which still rely on limited parallel data. The core bottleneck for open-ended idiomatic expression polishing is the lack of large-scale datasets with explicit supervision—a gap our work directly addresses by generating data with LTP annotations.

2.2 LLM-SLM Collaborative Mechanism

Collaboration between LLMs and SLMs has emerged as a promising paradigm to tackle data scarcity (Tan et al., 2024; Panchbhai and Pankanti, 2021; Chen et al., 2025). The typical synergy involves using an LLM for draft generation or knowledge expansion, and an SLM for domain-specific refinement or validation (Zhang et al., 2024), as seen in scientific literature analysis (Li et al., 2025), code optimization (Luo et al., 2025), clinical decision support (Bao et al., 2023), and legal text processing (Yue et al., 2023). Frameworks like FreeAL (Xiao et al., 2023) further explore using SLMs to guide LLMs annotation actively. Different from these works which often aim to filter or enrich existing data, our Auto-IDEA framework leverages the LLM-SLM loop for a more foundational goal: the unsupervised creation of a large-scale, reasoning-annotated dataset from scratch for a novel generation task.

3 Methodology

We instantiate the Chinese Idiom Polishing (CIP) task through **Auto-IDEA**. In Figure 2, Auto-IDEA

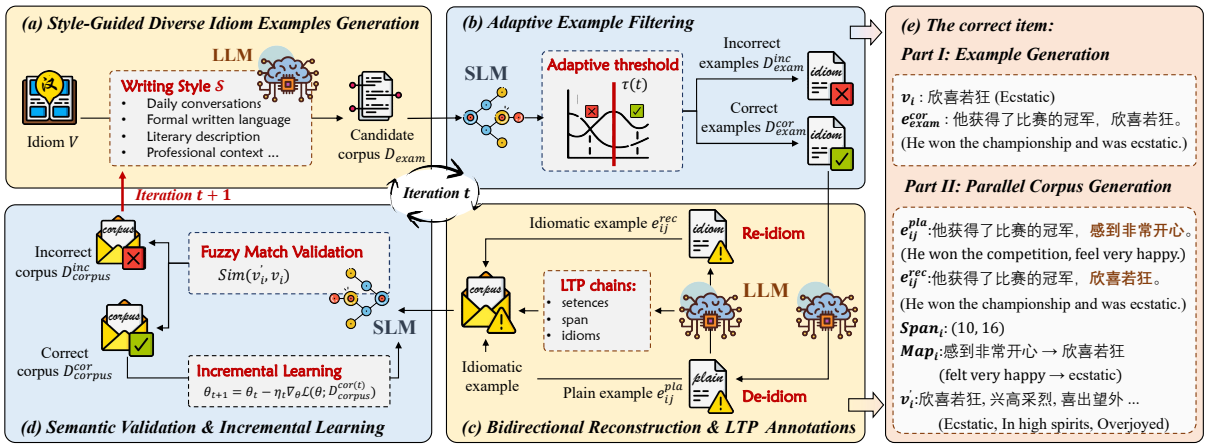

Figure 2: Overview of Auto-IDEA. In each loop, LLM (yellow) and SLM (blue) collaborate in four stages in sequence. (a) LLM generates idiom examples; (b) SLM filters correct expressions; (c) LLM constructs parallel corpus via semantic reconstruction; (d) SLM performs secondary filtering and enables incremental learning. Correct output forms a high-quality parallel corpus, while incorrect output (red arrows) triggers a regeneration generation loop. Subgraph (e) shows a parallel corpus item (orange).

operates through an iterative, four-stage collaboration between the LLM and SLM. Starting from a raw idiom lexicon, this closed-loop pipeline autonomously generates high-quality parallel corpora annotated with LTP chains by LLM, while simultaneously enabling the SLM to self-improve its validation capability.

3.1 Stage 1: Style-Guided Diverse Idiom Examples Generation

To overcome output homogeneity in conventional generation, we implement a style-guided prompting framework. Beginning with an idiom dictionary $\mathcal{V} = \{v_1, \ldots, v_m\}$, we define a style template $\mathcal{S} = \{s_1, \ldots, s_n\}$ covering diverse usage scenarios (as shown in Figure 2 (a)). For each idiom $v_i \in \mathcal{V}$, we generate multiple examples by associating each with a distinct style $s_j \in \mathcal{S}$.

Let ρ be a task-descriptive instruction to explain the generation objective, and e_{ij} denote the example generated for idiom v_i under style s_j. We define two core probability terms:

- $P_{ij}^{gen} = P(e_{ij} \mid v_i, s_j, \rho)$: the probability of generating the specific example e_{ij}.
- $P_j^{style} = P(s_j \mid s_{<j}, \mathcal{S})$: the probability of sampling style s_j given previously assigned styles, which enforces a uniform distribution without replacement.

The joint probability of generating the complete example set $\mathcal{E} = \{e_{ij}\}$ is therefore:

$$P(\mathcal{E} \mid \mathcal{V}, \mathcal{S}, \rho) = \prod_{i=1}^{|\mathcal{V}|} \prod_{j=1}^{|\mathcal{S}|} P_{ij}^{gen} \cdot P_j^{style} \quad (1)$$

where $\mathcal{E} = \{e_{ij} \mid 1 \le i \le |\mathcal{V}|, 1 \le j \le |\mathcal{S}|\}$ is the complete set of generated examples over idiom set $\mathcal{V}$.

The style sampling distribution $P_j^{style} = P(s_j \mid s_{<j}, \mathcal{S})$ ensures each style is used once per idiom:

$$P_j^{style} = \begin{cases} \frac{1}{|\mathcal{S}|-j+1} & \text{if } s_j \in \mathcal{S} \setminus \{s_1, \ldots, s_{j-1}\} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $\sum_{s_j \in \mathcal{S}} P_j^{style} = 1$. This mechanism prevents redundancy in the initial corpus $\mathcal{D}_{exam}$, and guarantees comprehensive coverage of all styles for each idiom. In Appendix A, we introduce the prompt engineering for generating idiom examples.

3.2 Stage 2: Adaptive Example Filtering

The incrementally trained SLM acts as a validator to filter the LLM-generated candidate examples $\mathcal{D}_{exam} = \{(e_{ij}, v_i)\}$ (see Figure 2 (b)). For each candidate pair (e_{ij}, v_i), the SLM performs binary validation, predicting a label $\hat{y}_{ij} \in \{+1, -1\}$ (where $+1$ denotes correct idiom usage, and -1 denotes incorrect usage). The decision is based on an adaptive threshold $\tau(t)$:

$$\hat{y}_{ij} = \begin{cases} +1 & \text{if } P(y = +1 \mid e_{ij}, v_i; \theta_t) \ge \tau(t), \\ -1 & \text{otherwise,} \end{cases} \quad (3)$$

where $P(y = +1 \mid e_{ij}, v_i; \theta_t)$ is the SLM's confidence score at iteration t with parameters θ_t. The threshold $\tau(t)$ decays exponentially across iterations:

$$\tau(t) = \tau_0 \cdot \gamma^t, \quad (\tau_0 = 0.85,\ \gamma = 0.9). \tag{4}$$

This decaying schedule ensures stringent filtering in early rounds and progressively relaxes the criterion as the SLM's validation capability improves through incremental learning.

The set of examples retained after filtering in iteration t is:

$$\mathcal{D}_{exam}^{cor} = \{(e_{ij}, v_i) \mid \hat{y}_{ij} = +1\}. \tag{5}$$

This curated set $\mathcal{D}_{exam}^{cor}$ proceeds to Stage 3 for parallel corpus construction.

3.3 Stage 3: Bidirectional Reconstruction and LTP Annotations

This stage takes the filtered idiom examples $\mathcal{D}_{exam}^{cor}$ as input and employs the LLM to perform bidirectional semantic reconstruction (see Figure 2 (c)). The core objective is to generate a high-quality parallel corpus where each pair is equipped with an explicit LTP annotation, providing the granular supervision signal required for training.

For each validated example $(e_{ij}, v_i) \in \mathcal{D}_{exam}^{cor}$, let $L(\cdot)$ denote the log-probability score from the LLM. The bidirectional reconstruction proceeds in two steps:

- **De-idiomatization**: The LLM generates a plain (idiom-free) version e_{ij}^{pla} from the idiomatic sentence e_{ij}. The generation follows: $P(e_{ij}^{pla} \mid e_{ij}) \propto \exp(L(e_{ij}^{pla} \mid e_{ij}, \phi_{free}))$, where ϕ_{free} is the instruction for generating text without idioms.

- **Re-idiomatization**: The LLM then reconstructs an idiomatic sentence e_{ij}^{rec} from the plain version e_{ij}^{pla}, conditioned on the original idiom's difficulty level $r(v_i)$: $P(e_{ij}^{\text{rec}} \mid e_{ij}^{pla}) \propto \exp(L(e_{ij}^{rec} \mid e_{ij}^{pla}, \phi_{r(v_i)}))$, where $\phi_{r(v_i)}$ is the instruction for generating text using idioms of difficulty $r(v_i)$.

The core innovation of this stage is the automatic derivation of the polishing location by comparing the three text versions. We algorithmically identify the contiguous span in e_{ij}^{pla} that differs from e_{ij}^{rec}, defining the location $Span_i$ as a tuple of start and end token indices (based on word segmentation):

$$Span_i = (\text{start\_idx}, \text{end\_idx}) \tag{6}$$

This index-based location $Span_i$, together with the rewrite idiom $v_i^{'}$, forms the complete LTP annotation $\text{LTP}_{ij} = (v_i^{'}, Span_i)$. It provides explicit, token-level supervision for *where* to edit.

The direct output of this stage is the fully annotated parallel corpus:

$$\mathcal{D}_{corpus} = \{(e_{ij}^{pla}, e_{ij}^{rec}, v_i, v_i^{'}, Span_i)\}. \tag{7}$$

The resulting LTP chain provides explicit supervision for the "locate-then-polish" mapping. For instance, Figure 1 illustrates how a sample teaches the model to locate the token span (12,17) and replace it with "欣喜若狂". This token-level localization granularity is fundamental for enabling autonomous learning of the LTP reasoning process. Figure 2 (e) shows the simplified data sample in CIP.

In Appendix A, we provide a detailed prompt instructions ϕ_{free} and $\phi_{r(v_i)}$.

3.4 Stage 4: Semantic Validation and Incremental Learning

This final stage implements the closed-loop feedback mechanism crucial to Auto-IDEA (see Figure 2 (d)). It serves two purposes: (1) validating the quality of the parallel corpus $\mathcal{D}_{corpus} = \{(e_{ij}^{pla}, e_{ij}^{rec}, v_i, v_i^{'}, Span_i)\}$ generated in Stage 3, where $v_i^{'}$ denotes the idiom used in the reconstruction, and (2) incrementally training the SLM on the validated data, thereby enhancing its ability to guide the entire framework in the next iteration.

We perform a two-step validation on each candidate tuple, using the original target idiom v_i from Stage 2 as the gold reference:

- **Exact Match**: The sample is accepted if the reconstructed idiom $v_i^{'}$ is identical to the original target v_i:

$$\mathcal{D}_{corpus}^{exact} = \{(e_{ij}^{pla}, e_{ij}^{rec}) \mid v_i^{'} = v_i\}. \tag{8}$$

- **Fuzzy Match**: For samples where $v_i^{'} \neq v_i$, we compute their semantic similarity using the SLM's embedding space:

$$\text{Sim}(v_i, v_i^{'}) = \cos\left(\mathbf{h}(v_i), \mathbf{h}(v_i^{'})\right), \tag{9}$$

where $\mathbf{h}(\cdot)$ is the embedding from the SLM. Samples with a similarity score above a threshold σ are retained:

$$\mathcal{D}_{corpus}^{fuzzy} = \{(e_{ij}^{pla}, e_{ij}^{rec}) \mid \text{Sim}(v_i, v_i^{'}) \geq \sigma\} \tag{10}$$

The final validated corpus combines these subsets:

$$\mathcal{D}_{corpus}^{cor} = \mathcal{D}_{corpus}^{exact} \cup \mathcal{D}_{corpus}^{fuzzy} \tag{11}$$

The validated examples $\mathcal{D}_{\text{corpus}}^{\text{cor}}$ in iteration t then drive incremental updates to the SLM parameters. The model optimization follows gradient descent:

$$\theta_{t+1} = \theta_t - \eta_t \nabla_\theta \mathcal{L}(\theta_t; \mathcal{D}_{\text{corpus}}^{\text{cor}(t)}) \tag{12}$$

where the learning rate decays exponentially across iterations as $\eta_t = \eta_0 \cdot \beta^t, \quad \beta = 0.95$, gradually reducing update magnitudes as model performance stabilizes.

Samples that fail both validation stages, denoted as $\mathcal{D}_{corpus}^{inc}$, are sent back to Stage 1 for regeneration, creating a self-correcting loop. This closed-loop mechanism enables continuous refinement: validated examples augment the corpus while simultaneously enhancing the SLM's validation capability through incremental learning. The updated SLM then provides more accurate filtering in subsequent iterations, establishing a virtuous cycle that progressively improves both dataset quality and model performance across multiple refinement passes.

4 Experiments

4.1 Task Definition

Chinese Idiom Polishing (CIP) is a generative task that refines a plain text into a more idiomatic and stylistically enhanced version by substituting appropriate spans with Chinese idioms. Formally, the task requires a model to: (1) **Locate** one or more candidate spans in plain text that are semantically compatible with idioms, and (2) **Polish** each span with a contextually and culturally appropriate idiom. This **Locate-Then-Polish** process distinguishes CIP from cloze-style idiom understanding, as it demands autonomous reasoning without prespecified candidates or positions.

4.2 The CIP-200K Dataset

We conduct experiments on **CIP-200K**, a large-scale dataset for Chinese Idiom Polishing generated automatically by the Auto-IDEA framework. It comprises 206,342 parallel sentence pairs, each annotated with an LTP chain. The dataset is split into Training (136,342 pairs), Development (20,000 pairs), and Test (50,000 pairs) sets (see Table 1). To rigorously evaluate generalization, 13% of idioms in the test set are completely unseen during training. To substantiate the reliability and richness of CIP-200K, we present a multi-faceted analysis covering its statistical properties, structural characteristics, and expert-validated quality in Appendix B.

Data Split	Train	Dev	Test
Sentence pairs	136,342	20,000	50,000
Idioms	31,784	2,099	5,039
Text Characteristics	All Dataset		
	Original	Polished	LTP
Avg. tokens	85.23	60.07	587.16
Avg. words	11.00	10.77	103.88

Table 1: Statistical properties of the CIP-200K dataset.

4.3 Baseline Models

We establish a comprehensive set of baseline models by systematically selecting state-of-the-art open-source and proprietary LLMs. For open-source models, we include three advanced architectures with comparable parameter scales: **Qwen3-8B**, **Llama3-8B** and **GLM4-9B**. For proprietary systems, we evaluate three leading commercial models: **DeepSeek-R1**, **GPT-4** and **Claude-4**. For specific experimental settings and parameters, please refer to Appendix C.

4.4 Evaluation Metric

We evaluate the task on the test set of CIP-200K. During the evaluation process, we set the temperature parameter $t = 0.3$, and limited the maximum output length to 512 tokens. Our evaluation objective is the quality of the final polished sentences. Besides using common evaluation metrics such as PPL, STS, BLEU-4 and ROUGE-L, we also propose two evaluation metrics for idiom polishing task as the main metrics: **Idiom Polishing Accuracy (IPA)** and **Text Compression Ratio (TCR)**:

$$\text{IPA} = \frac{\text{\# Correct Idioms}}{\text{\# Total Idioms}} \tag{13}$$

$$\text{TCR} = 1 - \frac{\ell_{\text{polished}}}{\ell_{\text{original}}} \tag{14}$$

Model	Method	CIP						ChID	CINLID
		IPA↑	TCR*	STS↑	PPL↓	BLEU-4↑	ROUGE-L↑	ACC↑	ACC↑
Llama3-8B	Base	0.1763	0.4176	0.8710	65.20	0.1945	0.0082	0.2601	0.6893
	SFT	0.2573	0.6877	0.6124	60.50	0.4052	0.0485	0.2735	0.6577
	LTP(ours)	0.4825	0.2424	0.9258	49.15	0.6124	0.3184	0.3017	0.7250
GLM4-9B	Base	0.3464	0.2279	0.8681	46.50	0.2032	0.0259	0.5220	0.7882
	SFT	0.3208	0.3251	0.8133	52.80	0.2165	0.3002	0.5323	0.7896
	LTP(ours)	0.5915	**0.3004**	0.9246	46.85	0.5587	**0.7253**	0.5998	0.8199
Qwen3-8B	Base	0.1484	0.2910	0.8643	49.50	0.3045	0.0652	0.5835	0.7087
	SFT	0.3899	0.2877	0.7860	58.50	0.1925	0.0805	0.5988	0.7554
	LTP(ours)	**0.6415**	0.2950	0.9321	39.32	0.7528	0.3854	0.6493	0.8396
Proprietary Models	GPT-4	0.5060	0.0235	**0.9873**	**39.52**	0.8743	0.2165	0.6177	0.7922
	Claude-4	0.5525	-0.0125	0.9784	38.47	**0.8845**	0.3725	**0.8367**	**0.8589**
	DeepSeek-R1	0.5791	-0.1253	0.9521	40.25	0.8643	0.2425	0.6859	0.8277

Table 2: Model performance on Chinese Idiom Polishing Task. Yellow/blue shadows denote core polishing (IPA↑, TCR↑*, STS↑) and generation metrics (PPL↓, BLEU-4↑, ROUGE-L↑) respectively. Orange indicates external benchmarks. *TCR ≈ 0.3 is optimal: values > 0.3 may cause semantic loss, while values < 0 indicate verbose degeneration. **Bold** and underlined mark best/suboptimal results.

where the best effect is achieved when TCR is approximately equal to 0.3 (refer to the Text Characteristics in Table 1). A TCR value that is too large may lead to the loss of key information, while a negative TCR value indicates that the polished sentence is lengthier.

5 Results and Analysis

5.1 Main Results

We conduct a comprehensive experiment on the CIP-200K dataset with LTP annotations. Table 2 reveals that standard supervised fine-tuning (SFT) on our generated data yields unstable and often detrimental effects: while it may improve certain metrics (e.g., Qwen3's IPA from 0.1484 to 0.3899), it frequently degrades semantic fidelity (e.g., STS drops for all models), and can even harm core idiom mastery (e.g., GLM4's IPA drops from 0.3464 to 0.3208).

In stark contrast, our LTP fine-tuned models successfully harnesses the valuable reasoning chains to overcome these limitations, consistently and substantially outperforming both the base model and the SFT baseline across every evaluation dimension for all open-source models. This synergy demonstrates (1) **the utility of the structured supervision in CIP- 200K**, and (2) **the superior effectiveness and robustness of the LTP fine-tuning paradigm over conventional SFT for idiom polishing**. We analyze the results from three primary dimensions.

(a) Core Idiom-Specific Capabilities (IPA, TCR, STS). Our LTP strategy yields substantial and consistent gains in core idiom mastery. Idiom Polishing Accuracy (IPA) improves dramatically over SFT (+22.5 points for Llama3, +27.1 for GLM4, +25.2 for Qwen3), with Qwen3-LTP (0.6415) surpassing GPT-4. For Text Compression Ratio (TCR), LTP achieves near-optimal balance ($\sim$0.3), avoiding the over-compression of SFT or the verbosity of some proprietary models. Crucially, LTP fully mitigates the semantic erosion seen with SFT, restoring Semantic Text Similarity (STS) to high levels ($>$0.92) for all models.

(b) Surface Generation Quality (PPL, BLEU-4, ROUGE-L). LTP also enhances overall text quality. It reduces Perplexity (PPL) substantially compared to SFT (e.g., Qwen3-LTP: 39.32 vs. SFT's 58.50), even outperforming GPT-4. The dramatic surge in BLEU-4 (Qwen3-LTP: 0.7528 vs. SFT's 0.1925) confirms precise idiom localization. Trends in ROUGE-L highlight model-specific strengths in coverage or coherence, contrasting with SFT's frequent degradation of fluency (GLM4-LTP: 0.7253 vs. SFT's 0.3002).

(c) Generalization to Downstream Tasks (ChID, CINLID). The learned knowledge transfers robustly. On ChID, Qwen3-LTP achieves 0.6493 accuracy (an 8.4-point gain over SFT). On CINLID, it attains 0.8396 accuracy, surpassing GPT-4 by 5.97 points and approaching the top-performing model. These gains confirm that LTP cultivates

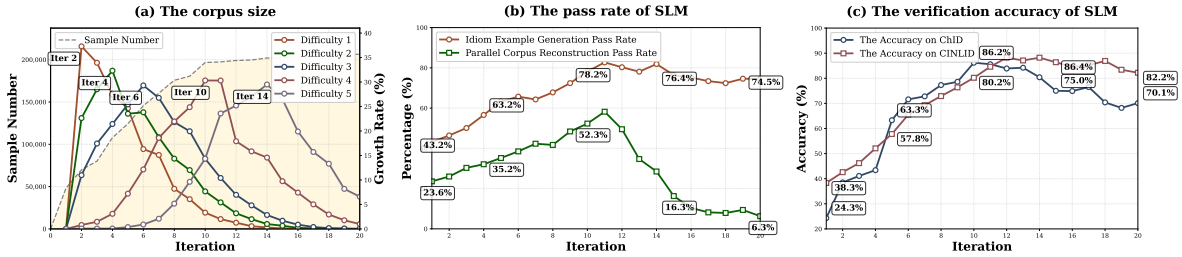

Figure 3: Performance analysis of the Auto-IDEA Framework during 20-round data generation, where the X-axis represents the number of iteration rounds (a total of 20 rounds). Subfigure (a) tracks the growth of the generated corpus size over training iterations; Subfigure (b) shows the pass rate of the Small Language Model (SLM) for two key generation stages; Subfigure (c) depicts the downstream task accuracy of the SLM over iterations.

deep, transferable comprehension of idiomatic semantics.

To validate that the automatic performance gains translate into perceived quality, we conducted a rigorous human evaluation comparing Qwen3-LTP against Qwen3-Base and the strong proprietary model DeepSeek (full protocol in Appendix D). Evaluators rated Qwen3-LTP highest in **Overall Quality** (4.34/5), and notably in the core dimensions of **Idiomaticity** (4.35) and **Conciseness** (4.28). This human judgment confirms that LTP fine-tuning yields not only superior automatic scores but also generates outputs that are perceived as more authentic and effective by native speakers.

5.2 In-depth Analysis

We analyze the internal dynamics of Auto-IDEA across 20 iterations to validate its self-improving nature. Tracking corpus growth, SLM validation pass rates, and the SLM's own downstream accuracy reveals a synergistic co-evolution of data quality and model capability.

(a) Auto-IDEA is an adaptive and scalable data generation framework. As shown in Figure 3 (a), the validated corpus size plateaus after iteration 10, reaching 206,342 samples. Crucially, lower-difficulty samples (Level 1-3) saturate early, while the peak growth rates for higher-difficulty samples (Level 4-5) are delayed, occurring sequentially around iterations 10 and 14. This demonstrates that SLM incremental learning effectively shifts the framework's focus toward increasingly challenging instances, enabling difficulty-aware corpus expansion without external intervention.

(b) The SLM's filtering mechanism facilitates generation of higher-difficulty, higher-quality data. Figure 3 (b) shows that the SLM's validation pass rates for both idiom generation and parallel corpus reconstruction increase through iteration 10, reflecting its enhanced filtering capability. Beyond this point, pass rates gradually decline as the proportion of high-difficulty samples rises. This indicates not a loss of competence, but a tightening of quality standards when assessing more complex cases—a key mechanism for maintaining data integrity while scaling difficulty.

Model	Stage 2	Stage 3	Stage 4	IPA (%)
Auto-IDEA	✓	✓	✓	45.68
Variant 1	×	✓	✓	41.96 (↓3.7)
Variant 2	✓	×	✓	32.12 (↓13.6)
Variant 3	✓	✓	×	43.03 (↓2.7)

Table 3: The Ablation Study of Auto-IDEA Framework. Variant 1 excludes the adaptive example filtering in Stage 2, Variant 2 removes the CoT construction in Stage 3, and Variant 3 disables the closed-loop semantic verification in Stage 4. ✓ = enabled, × = disabled.

(c) SLM's downstream performance improvement confirms enhanced idiom judgment capability. The SLM's own accuracy on external idiom benchmarks (ChID, CINLID) improves steadily, peaking at iteration 12 with final gains of +43.9% and +45.8%, respectively (Figure 3 (c)). The subsequent slight decline correlates with the increased challenge of late-stage validation samples. This trajectory confirms that the SLM, through incremental learning on curated data, develops robust idiomatic knowledge, directly translating into stronger downstream judgment—completing the virtuous cycle of data-quality and model-capability co-evolution.

5.3 Ablation Study

We conduct an ablation study to quantify each core component's contribution. Fine-tuning Qwen3-8B on a fixed 30K-sample subset from iteration 5 shows that removing any stage degrades IPA (Table 3), confirming the necessity of the full pipeline.

Original Sentence: 公司现在这状况真是快要垮了，你还在那儿优哉游哉地摸鱼，赶紧想想办法吧！

(The company is on the verge of collapse, and you're still idling away. Hurry up and find a solution!)

Base Model: 公司现在这状况真是快要垮了，你还在那儿悠然自得地摸鱼，赶紧想想办法吧！

(The company is on the verge of collapse, and you're still leisurely idling. Hurry up!)

LTP-Finetuned Model: 公司现在这状况真是大厦将倾，你还在那儿优哉游哉地摸鱼，赶紧想想办法吧！

(The company's state is like a great building about to collapse, and you're still idling away. Hurry up!)

Table 4: A case contrasting the base and LTP-finetuned Qwen3-8B, highlighting the critical improvement in *locating accuracy*.

(a) Critical Role of Chain-of-Thought Construction. The most severe drop (-13.6%) occurs in Variant 2, where bidirectional semantic reconstruction (Stage 3) is disabled. This confirms that the explicit LTP reasoning chains provide the essential supervisory signal for learning precise, position-aware editing.

(b) Importance of Iterative Collaboration. Performance also declines when removing adaptive filtering (Variant 1, -3.7%) or closed-loop validation (Variant 3, -2.7%). These results underscore the SLM's iterative role in filtering noise and refining semantic alignment, validating that LLM-SLM synergy is key to achieving high data quality.

5.4 Qualitative Case Analysis

This section provides a qualitative case analysis to illustrate the fundamental shift in reasoning process brought by the LTP fine-tuning paradigm. In contrast to the Base Model, the LTP-Finetuned Model demonstrates a precise ability to target genuine lexical redundancy while preserving contextually appropriate expressions, thereby effectively eliminating idiom misuse. As shown in Table 4, the base model incorrectly substitutes the negative expression "优哉游哉" (idling away negligently) with the positive idiom "悠然自得" (leisurely), introducing a stylistic mismatch. Conversely, the LTP-finetuned model first correctly locates the core issue—the colloquial phrase "快要垮了" (on the verge of collapse) as the most semantically redundant and stylistically weak span. It then selects the semantically precise idiom "大厦将倾" (like a great building about to collapse) through contextual metaphor mapping. This substitution perfectly preserves the original sense of urgency and negative connotation. For a systematic analysis covering more challenge types, please refer to Appendix E.

5.5 Generalization experiments in English idioms

To evaluate the **task-general applicability** of the Auto-IDEA framework, we apply it to the established English idiom processing benchmark—the EPIE Corpus (Saxena and Paul, 2020). Our objective is to test whether the framework can autonomously generate LTP chains for a **different task formulation** (span detection and idiomatization), leading to superior performance. We adapt Auto-IDEA to consume EPIE's idiom location tags and candidate lists, iteratively producing a novel dataset, **EPIE-LTP**, enriched with explicit reasoning chains for idiom identification and paraphrasing. Fine-tuning Llama3-8B, GLM4-9B, and Qwen3-8B on EPIE-LTP yields our models, collectively referred to as **LTP-EN**.

Table 5 presents the comprehensive results for the idiomatization task across the three LLMs. Our LTP-EN method demonstrates clear advantages in span localization while showing nuanced performance patterns in text generation.

Across all three models, LTP-EN consistently achieves the highest localization metrics, with particularly notable gains in F1-score (Llama3-8B: 0.8801 vs. SFT-EN's 0.6154; Qwen3-8B: 0.8374 vs. 0.5883). This confirms that explicit reasoning chains provide effective supervision for boundary detection, a critical requirement for idiomatization. The improvement is most pronounced for Llama3-8B, where LTP-EN shows a 43% relative improvement in F1 over SFT-EN.

While LTP-EN excels at localization, generation quality presents a more complex picture. For Llama3-8B, SFT-EN achieves higher BLEU-4 (0.6004) and ROUGE-L (0.6503) than LTP-EN (0.5312, 0.4875), suggesting a potential trade-off between precise localization and fluent generation. However, this pattern does not generalize across all models: Qwen3-8B's LTP-EN shows balanced improvements in both localization (F1: 0.8374) and generation (BLEU-4: 0.5742).

These results validate Auto-IDEA's effectiveness in enhancing structural understanding while highlighting model-specific responses to reasoning chain training. The framework proves particularly valuable for tasks requiring precise text manipulation, though optimal balancing between localiza-

Model	Method	Localization Metrics				Generation Metrics	
		ACC↑	Precision↑	Recall↑	F1↑	BLEU-4↑	ROUGE-L↑
Llama3-8B	Base-EN	0.6877	0.6541	0.5987	0.6251	0.3204	0.4805
	SFT-EN	0.6581	0.7845	0.6547	0.6154	**0.6004**	**0.6503**
	LTP-EN (ours)	**0.8613**	**0.8952**	**0.8677**	**0.8801**	0.5312	0.4875
GLM4-9B	Base-EN	0.5435	0.4563	0.6543	0.5432	0.2152	0.2656
	SFT-EN	0.5567	0.3782	0.6876	0.5913	0.3852	0.2357
	LTP-EN (ours)	0.6754	0.5214	0.7032	0.6831	0.4352	0.4652
Qwen3-8B	Base-EN	0.5103	0.6583	0.5914	0.4733	0.3255	0.4757
	SFT-EN	0.6884	0.5757	0.6023	0.5883	0.3959	0.3456
	LTP-EN (ours)	0.8322	0.8462	0.8283	0.8374	0.5742	0.5727

Table 5: Model performance on English Idiom Transformation Task (Idiomatization). Yellow/blue shadows denote localization and generation metrics respectively. The Idiomatization task requires models to identify literal expressions and replace them with appropriate idioms. Our LTP-EN model, trained with explicit reasoning chains from Auto-IDEA, consistently outperforms both Base-EN and SFT-EN across all metrics for all three LLMs. **Bold** indicates the best performance for each metric within each model family; underline marks suboptimal results where another model performs better overall.

tion and generation objectives may require model-specific adjustments.

6 Conclusion

We propose Auto-IDEA—a novel data generation paradigm that establishes cyclic collaboration between LLMs (executing bidirectional semantic reconstruction with style control for corpus diversity) and SLMs (performing incremental validation with Chain-of-Thought construction). This synergy generates CIP-200K: the first large-scale dataset with granular reasoning annotations, effectively eliminating human supervision while resolving data scarcity. Building on this, we convert the proposed Locate-Then-Polish (LTP) annotations into teachable reasoning patterns, achieving state-of-the-art performance (64.2% IPA on Qwen3-8B). Human evaluations further affirm both the quality of the generated data and the resulting polished outputs. More broadly, Auto-IDEA establishes an annotation-free, self-iterating, and scalable data generation framework that advances research in semantic refinement and low-resource language learning. Future work will extend its applicability to other languages while prioritizing implementation in educational contexts.

Acknowledgements

This work is supported in part by the Natural Science Foundation of China (grant No.62276188) and Original Exploration Program of the National Natural Science Foundation (grant No.62550068).

Limitations

The Auto-IDEA framework's effectiveness is subject to three core constraints. First, the quality and diversity of its self-generated corpus (e.g., CIP-200K) are intrinsically bounded by the knowledge and biases of the initial LLM, which may underrepresent complex or rare idiomatic expressions. Second, while the framework shows promise in English idiom detection, its generalizability to generative idiomatic tasks in other languages with different linguistic structures remains unverified. Finally, the current "Locate-Then-Polish" paradigm struggles with idioms requiring deep cultural allusions or nuanced discourse-level reasoning, indicating a ceiling for fully automated handling of high-complexity language phenomena. During the preparation of this work, we used an AI tool for language polishing and proofreading. After using this tool, we reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *ArXiv*, abs/2308.14346.

Yi Chen, JiaHao Zhao, and HaoHao Han. 2025. A survey on collaborative mechanisms between large and small language models. *Preprint*, arXiv:2505.07460.

Ertürk ERDAĞI. 2024. Use of natural language processing methods in teaching turkish proverbs and idioms. *International Journal of Advanced Computer Science & Applications*, 15(7).

Sebastián Ismael Gárate Rodas and Andrea Katherine Palacios Alvarado. 2025. Non-native english speakers' strategies for learning idiomatic expressions.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. 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 18554–18563. AAAI Press.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2025. Scilitllm: How to adapt llms for scientific literature understanding. *Preprint*, arXiv:2408.15545.

Siyu Long, Ran Wang, Kun Tao, Jiali Zeng, and Xinyu Dai. 2020. Synonym knowledge enhanced reader for Chinese idiom reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3684–3695, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2025. Wizardcoder: Empowering code large language models with evol-instruct. *Preprint*, arXiv:2306.08568.

Anand Panchbhai and Smarana Pankanti. 2021. Exploring large language models in a limited resource scenario. In *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, pages 147–152.

Byron Rene Pintado and Tammy Fajardo. 2021. Learning idioms through the multimodal approach. *Journal of Education and Practice*.

Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2023. Chinese idiom paraphrasing. *Transactions of the Association for Computational Linguistics*, 11:740–754.

Narutatsu Ri, Bill Sun, Sam Davidson, and Zhou Yu. 2023. Ideals: Idiomatic expressions for advancement of language skills. *arXiv preprint arXiv:2305.13637*.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions. In *International Conference on Text, Speech, and Dialogue*, pages 87–94. Springer.

Ying Sha, Mingmin Wu, Zhi Zeng, Xing Ge, Zhongqiang Huang, and Huan Wang. 2023. A prompt-based representation individual enhancement method for chinese idiom reading comprehension. In *Database Systems for Advanced Applications - 28th International Conference, DASFAA 2023, Tianjin, China, April 17-20, 2023, Proceedings, Part III*, volume 13945 of *Lecture Notes in Computer Science*, pages 682–698. Springer.

Yutong Shao, Rico Sennrich, Bonnie Webber, and Federico Fancellu. 2018. Evaluating machine translation performance on Chinese idioms with a blacklist method. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Minghuan Tan and Jing Jiang. 2021. Learning and evaluating Chinese idiom embeddings. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1387–1396, Held Online. INCOMA Ltd.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Tongguan Wang, Mingmin Wu, Guixin Su, Dongyu Su, Yuxue Hu, Zhongqiang Huang, and Ying Sha. 2025. Mchirc: A multimodal benchmark for chinese idiom reading comprehension. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 25398–25406. AAAI Press.

Xiaoyang Wang, Chen Li, Jianqiao Zhao, and Dong Yu. 2021. Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14006–14014. AAAI Press.

Lung-Hsiang Wong, Chee-Kuen Chin, Chee-Lay Tan, May Liu, and Cheng Gong. 2010. Students' meaning making in a mobile assisted chinese idiom learning environment. In *Learning in the Disciplines: Proceedings of the 9th International Conference of the Learning Sciences, ICLS 2010, Chicago, IL, USA, June 29 - July 2, 2010, Volume 1*, pages 349–356. International Society of the Learning Sciences / ACM DL.

Mingmin Wu, Guixin Su, Yongcheng Zhang, Zhongqiang Huang, and Ying Sha. 2024. Refining idioms semantics comprehension via contrastive

learning and cross-attention. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13785–13795, Torino, Italia. ELRA and ICCL.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. FreeAL: Towards human-free active learning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14535, Singapore. Association for Computational Linguistics.

Shengbin Yue, Wei Chen, Siyuan Wang, Bingxuan Li, Chenchen Shen, Shujun Liu, Yuxuan Zhou, Yao Xiao, Song Yun, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-lawllm: Fine-tuning large language models for intelligent legal services. *Preprint*, arXiv:2309.11325.

Yunisa Zahara and Heni Ekawati. 2025. Culturally aware and adapted nlp: Towards inclusive language learning tools. *NeuroLingua: Journal of Cognitive, Technological, and Cultural Language Learning*, 1(1):12–20.

Jing Zhang, Hui Gao, Peng Zhang, Boda Feng, Wenmin Deng, and Yuexian Hou. 2024. La-ucl: Llm-augmented unsupervised contrastive learning framework for few-shot text classification. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pages 10198–10207.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy. Association for Computational Linguistics.

A Prompt Engineering for Auto-IDEA

This chapter takes the Chinese idioms polishing as an example to introduce in detail the prompt engineering for each stage of Auto-IDEA and the generated content. **Prompt A** corresponds to instrction ρ in Stage 1 (Section 3.1), and **Prompt B** and **Prompt C** respectively correspond to ϕ_{free} and $\phi_{r(v_i)}$ in Stage 3 (Section 3.3). For readability, all prompts are presented in English in this paper, though Chinese was used in the actual experiments.

B The Dataset Analysis and Validation

In this section, we will provide a detailed introduction to the data sources, data analysis, and quality assessment of the CIP-200K dataset.

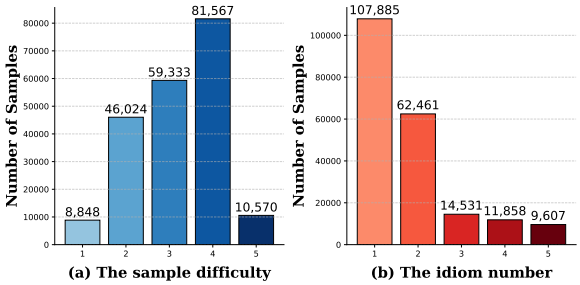

Figure 4: Distribution of sample data.

B.1 Data Source

We constructed a lexicon comprising 32,451 idioms, which were collected through the following source: the ChID dataset[1], the CINLID dataset[2], the CCT dataset[3], the Xinhua Dictionary[4] and the Idiom Encyclopedia[5]. The idiom lexicon be used as the input for Auto-IDEA to start the Chinese idiom polishing task.

B.2 Corpus Statistics and Difficulty Distribution

The foundational statistics in Table 1 confirm the corpus's large scale and the substantial lexical compression achieved by idiom polishing. Crucially, the dataset is annotated with a five-level difficulty score (1: Very Easy to 5: Very Hard), assigned by the SLM during generation based on idiom rarity and contextual complexity.

Figure 4 (a) reveals the distribution of this difficulty. It follows a bell-shaped curve centered on Levels 3 and 4, indicating that the Auto-IDEA framework most naturally generates samples of moderate complexity—a reflection of common idiomatic usage. The lower proportions of Level-1 (very simple) and Level-5 (very complex) samples highlight the framework's inherent challenge in generating trivial or extremely difficult examples, making the existing high-difficulty samples particularly valuable for stress-testing models.

Figure 4 (b) shows the distribution of the number of idioms per polished sentence. While most sentences (over 75%) contain one or two idioms, a significant portion (approx. 15%) incorporates three or more, with some containing up to five. This distribution confirms that CIP-200K captures a wide spectrum of polishing complexity, from sin-

[1] https://github.com/chujiezheng/ChID-Dataset
[2] https://openi.pcl.ac.cn/ZhangbuDong/CINLID
[3] https://github.com/bazingagin/chengyu_data
[4] https://github.com/pwxcoo/chinese-xinhua
[5] http://www.guoxue.com/

gle substitutions to dense, multi-idiom rewrites, posing a non-trivial generation challenge.

B.3 Human Evaluation of Data Fidelity

To provide expert-validated evidence for the quality of the CIP-200K dataset, we conducted a rigorous and fine-grained human evaluation. This evaluation aims to assess whether the automatically generated sentence pairs and their LTP chains meet the high standards required for a reliable benchmark.

The evaluation was carried out by five independent annotators, all of whom are native Chinese speakers with postgraduate degrees in linguistics or computational linguistics. Each annotator underwent a training session with detailed guidelines and practice samples to ensure a consistent understanding of the evaluation criteria.

We randomly sampled 300 instances from the CIP-200K, ensuring coverage across all five difficulty levels. Each annotator evaluated all 300 samples. For each sample, they were presented with the original plain sentence S_{plain}, the polished sentence S_{idiom}, and the corresponding LTP chain. They were asked to rate the sample on five distinct dimensions using a 5-point Likert scale (1: Very Poor, 2: Poor, 3: Acceptable, 4: Good, 5: Excellent). Our evaluation framework is designed to holistically assess the quality of CIP-200K:

- **Semantic Faithfulness**: Does the polished sentence S_{idiom} preserve the core factual meaning and intent of the original sentence S_{plain}?
- **Idiomatic Appropriateness**: Is the substituted idiom the optimal choice for the identified span in the given context?
- **Grammaticality & Fluency**: Is the polished sentence S_{idiom} grammatically correct and naturally fluent in modern Chinese?
- **Coherence of Reasoning Chain**: Is the provided LTP chain logically coherent and clearly justified?
- **Overall Quality**: Considering all factors above, what is the holistic quality of this data sample as an instance of the CIP task?

The results of the human evaluation are summarized in Table 6. The high average scores (ranging from 4.28 to 4.51 across dimensions) provide strong evidence for the high quality of the CIP-200K dataset. The Overall Quality score of 4.43 indicates that the samples are, on average, judged to be between "Good" and "Excellent". The slightly lower score for Coherence of Reasoning Chain (4.28) indicates that some non-core words may have a negative impact on the span determination. This will be explained in detail in the subsequent case analysis. Crucially, the high scores in Semantic Faithfulness (4.51) and Idiomatic Appropriateness (4.35) confirm that the core task—accurately locating and substituting with a contextually correct idiom—is successfully accomplished by the Auto-IDEA framework.

The substantial inter-annotator agreement, measured by Fleiss'Kappa (κ), ranges from 0.72 to 0.85, indicating "substantial" to "almost perfect" agreement according to common interpretation scales. This high consistency underscores the reliability of both the evaluation protocol and the inherent quality of the annotated data.

C Training Configuration

We perform supervised fine-tuning (SFT) directly on the CIP-200K dataset, where each training instance consists of a plain sentence paired with its idiomatic counterpart and the corresponding Locate-Then-Polish (LTP) reasoning chain. This allows models to jointly learn the mapping and the underlying locate-and-substitute rationale in a single stage. All models are trained with identical hyperparameters for 3 epochs, employing BF16 mixed-precision training with a learning rate of 2×10^{-6} (with 10% warmup steps), and a maximum sequence length of 1024. All experiments were conducted on 2 $\times$ A100 40 GB GPUs using data parallelism, with each device processing 2 samples per batch.

D Human Evaluation Protocol and Results

D.1 Experimental Design

D.1.1 Annotator Recruitment and Training

We recruited three expert annotators, all native Mandarin Chinese speakers with postgraduate degrees in linguistics or Chinese language and literature. Prior to the evaluation, all annotators underwent a standardized 1-hour training session that included:

Evaluation Dimension	Avg. Score (1–5)	Std. Dev.	Fleiss' Kappa (κ)
1. Semantic Faithfulness	4.51	0.62	0.81
2. Idiomatic Appropriateness	4.35	0.71	0.77
3. Grammaticality & Fluency	4.61	0.55	0.85
4. Coherence of Reasoning Chain	4.28	0.75	0.72
5. Overall Quality	4.43	0.66	0.79

Table 6: Human Evaluation Results on Data Fidelity ($N = 300$ samples).

- An introduction to the Chinese Idiom Polishing task and its challenges.
- Detailed explanation of the five evaluation dimensions and scoring criteria.
- Practice scoring on 20 sample outputs followed by discussion to calibrate understanding.

Annotators were compensated at standard academic rates.

D.1.2 Evaluation Protocol

We compared three systems: **Qwen3-Base**, **Qwen3-LTP** (our model), and **DeepSeek-R1**. From the CIP-200K test set, 200 instances were randomly sampled with stratification by idiom difficulty (20% from each of five difficulty levels). Each instance consisted of one original sentence and three polished versions (one from each model), presented in random order with model identities blinded.

Annotators used the detailed guideline shown in Tables 8 and 9 to score each polished sentence independently on five dimensions using a 5-point Likert scale (1: Poor, 5: Excellent). They were required to complete scoring for all three polished sentences before proceeding to the next set.

D.2 Results and Analysis

D.2.1 Summary Results

The human evaluation results are summarized in Table 7. Qwen3-LTP achieves the highest scores in *Idiomaticity* (4.35/5) and *Conciseness* (4.28), which are the most critical dimensions for the idiom polishing task. While DeepSeek-R1 performs well in *Semantic Faithfulness* and *Fluency*, likely due to a conservative, literal rewriting strategy, Qwen3-LTP attains the highest *Overall Quality* score (4.34), demonstrating a superior balance between linguistic authenticity and communicative efficiency.

Model	Idiom.	Conc.	Sem.	Flu.	OvQ
Qwen3-Base	2.31	2.87	3.12	3.45	2.89
DeepSeek-R1	4.18	3.95	**4.23**	**4.32**	4.31
Qwen3-LTP	**4.35**	**4.28**	4.21	4.13	**4.34**

Table 7: Human evaluation scores.

D.2.2 Detailed Analysis

Several patterns emerge from the detailed scoring:

- **Idiomaticity-Conciseness Trade-off**: While DeepSeek achieves high fluency, annotators noted it occasionally forces unnatural idioms (described as "overuse"), whereas Qwen3-LTP shows better judgment in idiom selection and integration.
- **Semantic-Rhetorical Balance**: DeepSeek's marginally higher Semantic score reflects a conservative, literal rewriting strategy that minimizes meaning change but may miss idiomatic opportunities. Qwen3-LTP achieves comparable semantic fidelity while being more rhetorically effective.
- **Overall Quality Consensus**: Despite minor variations in individual dimensions, all three annotators consistently rated Qwen3-LTP highest in Overall Quality, indicating a clear holistic preference for our model's outputs.
- **Base Model Limitations**: Qwen3-Base scored significantly lower across all dimensions, particularly in Idiomaticity (2.31), confirming that without specialized training, base LLMs struggle with culturally-grounded idiom polishing.

D.2.3 Reliability Analysis

Inter-rater reliability was calculated using the Intraclass Correlation Coefficient (ICC) for a two-way mixed-effects model assessing consistency

(ICC(3,k)). All five dimensions showed good to excellent reliability: Idiomaticity (ICC = 0.78), Conciseness (ICC = 0.71), Semantic Faithfulness (ICC = 0.83), Fluency (ICC = 0.76), and Overall Quality (ICC = 0.79).

E Extended Qualitative Analysis: Successes and Limitations

This appendix provides a comprehensive qualitative analysis of the LTP-finetuned model's performance, systematically comparing it with the base model across multiple dimensions. We present both *success cases* that demonstrate the clear advantages brought by the LTP fine-tuning paradigm and *optimizable cases* that reveal its current limitations and point to future research directions.

E.1 Success Cases: Demonstrating the Advantages of LTP Fine-tuning

The success cases in Table 10 vividly illustrate how the LTP fine-tuning paradigm fundamentally improves the model's capability to handle the core challenges of the Chinese Idiom Polishing (CIP) task. Unlike the base model that often makes surface-level substitutions, the LTP-finetuned model demonstrates a deeper understanding of semantic salience, contextual appropriateness, and compositional coherence.

The cases in Table 10 collectively demonstrate that the LTP paradigm enables the model to perform *localization-focused* and *context-aware* idiom polishing. By learning from explicit "Locate-Then-Polish" reasoning chains, the model acquires the ability to identify the most semantically redundant spans (Aspect 1), select idioms that match subtle contextual tones (Aspect 2), coordinate multiple substitutions for coherent rewriting (Aspect 3), and accurately address implicit rather than surface meanings (Aspect 4). These improvements directly explain the significant gains observed in quantitative metrics such as Idiom Polishing Accuracy (IPA) and Semantic Text Similarity (STS).

E.2 Optimizable Cases: Analyzing the Remaining Challenges

Despite significant improvements, by referring to the results of human evaluations, we found that there is still room for optimization in the LTP-finetuned model. Analyzing its persistent errors provides valuable insights into the inherent complexity of the CIP task and highlights promising directions for future research. Table 11 categorizes and analyzes four representative types of errors.

The error analysis in Table 11 reveals that the CIP task, at its highest level, requires knowledge and reasoning capabilities that extend beyond the current scope of our LTP framework. The remaining challenges involve **cultural and metaphorical reasoning** (Type A), **fine-grained semantic composition** (Type B), **discourse-aware coherence** (Type C), and **handling data imbalance for rare idioms** (Type D). These limitations are not unique to our approach but represent fundamental frontiers for research on idiomatic language processing. They suggest promising future work that could integrate external knowledge sources, more structured semantic representations, cross-sentence context modeling, and better data curation strategies.

F Task Generaliation to English Idiom Processing

This appendix details the application of the Auto-IDEA framework to the English Possible Idiomatic Expressions (EPIE) Corpus for the **idiomatization task**—identifying literal expressions and replacing them with appropriate idioms. We evaluate three LLMs to demonstrate the framework's broad applicability across different model architectures.

F.1 Experimental Setup

F.1.1 Task and Dataset

We use the EPIE dataset's idiom list and span annotations to construct the task data. Given a sentence containing a literal expression S_l, identify the span of L and generate a new sentence S_o where L is replaced by its corresponding idiom I. This task requires both precise localization and semantic transformation capabilities.

F.1.2 Models and Training

We evaluate three base models, each with three training approaches:

- **Base-EN**: The base model without any fine-tuning.
- **SFT-EN**: The model fine-tuned only on $\langle S_l, S_o \rangle$ pairs.
- **LTP-EN (Ours)**: The model fine-tuned on $\langle S_l, L, S_o \rangle$ triplets generated by Auto-IDEA.

All models are fine-tuned using LoRA with rank 16, learning rate 2e-4, batch size 32, and 5 epochs of training.

F.1.3 Evaluation Metrics

We evaluate two aspects of performance:

- **Localization Metrics**: Token-level Accuracy, Precision, Recall, and F1-score for identifying the target literal span in S_l.
- **Generation Metrics**: BLEU-4 and ROUGE-L scores for evaluating the quality of the generated idiomatic sentence S'_o compared to the gold reference S_o.

Prompt A: Example Generation for Chinese Idioms

Role: *system*
Content: You are an expert in Chinese creative writing and linguistics, proficient at generating diverse and contextually appropriate examples of idiom usage.
Role: *user*
Content: Please generate a unique and natural example sentence for the idiom `{idiom}` according to the following requirements:

- Sentence length should be between 30 and 70 characters.
- The target idiom `{idiom}` must be seamlessly incorporated. Additional idioms may be used if appropriate.
- Select one style from the **STYLE TEMPLATE** below and generate the example accordingly.
- Ensure originality and avoid common or stereotypical example sentences.
- Do not provide explanations or definitions of the idiom.

STYLE TEMPLATE:

1. *Casual conversation style* – using colloquial expressions.
2. *Formal written style* – suitable for news reports or academic writing.
3. *Literary descriptive style* – rich in imagery and rhetorical devices.
4. *Professional context* – in business or technology fields.
5. *Historical or traditional cultural context* – employing allusions or classical references.

Prompt B: Idiom-Free Rewriting

Role: *system*
Content: You are a professional Chinese NLP assistant specializing in text transformation and linguistic simplification.
Role: *user*
Content: Rewrite the `{sentence}` into a plain expression without using any idioms according to the following requirement:

- Replace all idioms while preserving the original semantic meaning.
- In the rewritten sentence, enclose each replaced idiom with # markers.
- Output only the rewritten sentence without any additional explanations or prefixes.

Prompt C: Idiom Replacement with Target Difficulty

Role: *system*
Content: You are a professional Chinese NLP assistant specializing in idiom replacement and text refinement.
Role: *user*
Content: Rewrite the following sentence by incorporating idioms of the specified difficulty level:
Original sentence: `{sentence}`
Difficulty level: `{difficulty}` (ranging from 1 to 5)
Requirements:

- Replace only the segments marked with # in the original sentence, maintaining the same #-marking for modified segments.
- Preserve the original sentence meaning and structure while replacing the marked segments.
- Use idioms with a difficulty level matching the target (1-5), where higher values indicate greater complexity or obscurity. Specific scoring criteria are defined in the reference guideline (Prompt D).

Prompt D: Scoring Criteria for the Difficulty of Idioms

Role: system
Content: You are a professional Chinese linguist. Please evaluate the difficulty of idioms according to the marking criteria.
Role: user
Content: Please assess the difficulty of the `{idiom}`. All scores are on an integer scale of 1 to 5 (from easy to difficult). The specific criteria are as follows:

1. Character Complexity

1. All characters are elementary-level Chinese characters
2. Contains 1 intermediate-level character or 2-3 structurally complex characters
3. Contains 1 rare character or multiple structurally complex characters
4. Contains 2 or more rare characters
5. Contains archaic or variant characters

2. Semantic Transparency

1. Literal meaning matches actual meaning
2. Requires simple metaphorical association
3. Literal/actual meanings are related but require explanation
4. Weak connection between literal/actual meanings
5. No apparent connection between literal/actual meanings

3. Cultural Background Depth

1. No specific cultural context required
2. Requires basic life knowledge
3. Requires historical/literary knowledge
4. Requires classical text knowledge
5. Requires obscure allusion knowledge

4. Modern Usage Frequency

1. High-frequency in daily speec
2. Common in formal writing
3. Domain-specific usage
4. Occasionally appears in literature
5. Nearly obsolete

Overall Score = 0.2 * Character + 0.3 * Semantic + 0.3 * Cultural + 0.2 * Frequency. Please provide your evaluation result:

HUMAN EVALUATION ANNOTATOR GUIDELINE	
1. Task Introduction	
Objective	Evaluate and compare the quality of **Chinese Idiom Polishing** from different AI systems. For each task instance, you will see one **Original Sentence** and three **Polished Sentences** (labeled A, B, C) generated by different models in a randomized and blinded order. Your task is to independently score each polished sentence across five dimensions.
Task Instance	• Original: [The plain input sentence] • Polished A: [First polished version] • Polished B: [Second polished version] • Polished C: [Third polished version]
Your Task	1. Read the Original sentence to understand its meaning. 2. Read each Polished sentence (A, B, C) carefully. 3. For **each** Polished sentence, assign a score (1-5) over **all five dimensions** listed in Table 9. 4. Scores are assigned independently per sentence. Complete scoring for all three polished sentences before proceeding to the next set.
2. Calibration Examples (Training)	
Example Set 1	*Original:* 他做事很小心，一点风险都不想有。(He does things very carefully and doesn't want to take any risks at all.) *Polished A:* 他做事如履薄冰，一点风险都不想有。(He acts with great caution, not wanting to take any risk at all.) **Rationale:** This is a **good** polishing example. The idiom "如履薄冰" (walking on thin ice) perfectly captures the meaning of being extremely cautious and risk-averse. It is contextually appropriate and feels natural. **Scores:** Idiom.=5, Conc.=4, Sem.=5, Flu.=5, OvQ.=5 *Polished B:* 他做事谨小慎微，规避风险。(He is cautious and risk-averse in his actions.) **Rationale:** This is a **medium** example. The idiom "谨小慎微" (cautious and meticulous) is appropriate, but "规避风险" is not an idiom and the original expression should be maintained. **Scores:** Idiom.=4, Conc.=2, Sem.=4, Flu.=5, OvQ.=3 *Polished C:* 他做事非常小心谨慎，一点风险都不想有。(He is extremely cautious in doing things and doesn't want to take any risks at all.) **Rationale:** This is a **poor** example. No idiom is used, making it fail the core requirement of idiom polishing. It simply repeats the original with minor rephrasing. **Scores:** Idiom.=1, Conc.=3, Sem.=5, Flu.=5, OvQ.=1
Example Set 2	*Original:* 这个计划考虑得不周全，有很多漏洞。(This plan is not well thought out and has many loopholes.) *Polished A:* 这个计划百密一疏，有很多漏洞。(This plan is full of flaws and loopholes.) **Rationale:** This is a **good** example. The idiom "百密一疏" (despite all precautions, there's still one oversight) accurately describes a plan that is mostly thorough but has a few flaws. It's contextually apt. **Scores:** Idiom.=5, Conc.=4, Sem.=5, Flu.=5, OvQ.=5 *Polished B:* 这个计划考虑得不全面，漏洞百出。(This plan is not comprehensive and full of loopholes.) **Rationale:** This is a **medium** example. While it uses the idiom "漏洞百出" (full of loopholes), it's partially redundant with "考虑得不全面." The polishing could be more concise. **Scores:** Idiom.=3, Conc.=3, Sem.=4, Flu.=5, OvQ.=3 *Polished C:* 这个计划设计有缺陷，存在许多问题。(This plan is flawed in design and has many problems.) **Rationale:** This is a **poor** example. It fails to use an appropriate idiom, using generic language instead. While semantically faithful and fluent, it doesn't achieve idiom polishing. **Scores:** Idiom.=1, Conc.=3, Sem.=4, Flu.=5, OvQ.=1

Table 8: Complete annotator guideline: Task introduction and calibration examples.

SCORING DIMENSIONS & CRITERIA (1=Poor, 5=Excellent)
1.Idiomaticity
1 (Poor): Idiom is completely misused, unnatural, or jarring in the context. The idiom does not fit the situation at all.
2 (Fair): Idiom usage is awkward and feels forced. The idiom is only marginally related to the context.
3 (Average): Idiom fits the context reasonably well but feels slightly unnatural or suboptimal. The meaning is conveyed but without elegance.
4 (Good): Idiom is well-chosen and appropriately used. It enhances the expression and feels mostly natural.
5 (Excellent): Idiom is flawlessly integrated, perfectly matching the context and significantly enhancing the expression. Usage feels completely natural and skilled.
2. Conciseness
1 (Poor): Sentence is wordy, repetitive, or longer than the original without adding value. Significant redundancy exists.
2 (Fair): Sentence is somewhat verbose with noticeable filler words or redundant expressions. Could be shortened without losing meaning.
3 (Average): Sentence length is acceptable but not optimal. Some minor redundancy remains.
4 (Good): Sentence is clearly more concise than the original. Efficiently conveys the same information with minimal words.
5 (Excellent): Sentence is crisp, efficient, and perfectly concise. Maximum information density achieved without sacrificing clarity.
3. Semantic Faithfulness
1 (Poor): Core meaning is altered, distorted, or completely lost. The polished sentence conveys a different message.
2 (Fair): Significant meaning changes or omissions. Key aspects of the original are missing or misrepresented.
3 (Average): Most of the original meaning is preserved, but with minor additions, omissions, or slight distortions.
4 (Good): Original meaning is well-preserved with only negligible changes. All key information is retained.
5 (Excellent): Perfectly preserves the original message, intent, and nuances. No meaning is lost or altered.
4. Fluency
1 (Poor): Sentence is ungrammatical, very choppy, or difficult to understand. Serious language errors present.
2 (Fair): Sentence has multiple grammatical errors or awkward phrasing that disrupts reading flow.
3 (Average): Sentence is understandable but contains some awkward expressions or minor grammatical issues.
4 (Good): Sentence reads smoothly with only minor, negligible issues. Flows naturally like native writing.
5 (Excellent): Sentence is grammatically perfect, coherent, and reads like well-written native text. No language errors.
5. Overall Quality
1 (Poor): Poor polishing that makes the sentence worse than the original. Multiple serious flaws across dimensions.
2 (Fair): Marginal improvement over original, but with significant flaws in multiple dimensions.
3 (Average): Adequate polishing with noticeable but not critical flaws. Represents a clear improvement over the original.
4 (Good): High-quality polishing with only minor weaknesses. Effectively achieves the goal of idiomatic refinement.
5 (Excellent): Excellent polishing that is a clear, effective improvement. Demonstrates skilled idiomatic rewriting.

Table 9: Scoring dimensions and criteria for annotator guideline.

Case Comparison between Base and LTP model

1. Precise Locating

• **Original:** 他在紧要关头迅速做出了选择，没有犹豫。(He quickly made a choice under time pressure.)
• **Base Model:**他在千钧一发的情况下迅速做出了选择，没有犹豫。 (He quickly made a choice under a critical situation, without hesitation.)
Issue: Misplaces focus on the circumstance rather than the core action is not the part that needs the most polishing.
• **LTP Model:** 他在时间紧迫的情况下当机立断，没有犹豫。 (He made a prompt decision under time pressure, without hesitation.)
Reasoning: Correctly locates the core action phrase "迅速做出了选择" and selects the idiom "当机立断", which precisely captures the meaning of making a quick and firm decision in a time-sensitive situation.

2. Contextual Appropriateness

• **Original:** 在激烈的辩论中，他坚持自己的看法，毫不退让。(In a heated debate, he insisted on his own view and did not yield.)
• **Base Model:** 在激烈的辩论中，他据理力争，毫不退让。(In the heated debate, he argued on the basis of reason, and did not yield.)
Issue: Semantically acceptable but fails to capture the confrontational tone and decisiveness implied by the context.
• **LTP Model:** 在激烈的辩论中，他力排众议，毫不退让。(In the heated debate, he overruled all objections, and did not yield.)
Reasoning: Correctly identifies that the phrase "坚持自己的看法，毫不退让" conveys strong insistence against opposition. The idiom "力排众议" (to overrule all objections) is selected as it precisely matches the scenario of standing firm against dissenting opinions in a confrontational setting, better capturing the contextual nuance than the more generic "据理力争".

3. Multi-Idiom Polishing

• **Original:** 他听到消息后非常震惊，愣在原地，然后才慢慢反应过来。(After hearing the news, he was extremely shocked, stood there dumbfounded, and then slowly came to his senses.)
• **Base Model:** 他听到消息后大惊失色，愣在原地，然后才慢慢反应过来。(After hearing the news, he was greatly alarmed and turned pale, stood there dumbfounded, and then slowly came to his senses.)
Issue: Only locates and polishes the first salient expression ("非常震惊"), leaving the other two ("愣在原地", "慢慢反应过来") unpolished. This reveals its inability to identify and coordinate multiple polishable spans within a single sentence.
• **LTP Model:** 他听到消息后愕然失色，呆若木鸡，良久才如梦初醒。(After hearing the news, he was stunned and pale, stood there like a wooden chicken, and only after a long while awoke as if from a dream.)
Reasoning: The LTP model is trained to identify *all* polishable spans. It successfully locates three consecutive spans that can be elevated: 1) "非常震惊", 2) "愣在原地", and 3) "慢慢反应过来". It then performs coordinated substitution with three semantically appropriate idioms, demonstrating its capacity for holistic multi-point polishing.

4. Richness of expression

• **Original:** 这份报告内容空洞，只是堆砌词藻。(This report is vacuous, merely piling up ornate words.)
• **Base Model:** 这份报告华而不实，只是堆砌词藻。(This report is flashy without substance, merely piling up words.)
Issue: Creates semantic redundancy because "华而不实" (gaudy but impractical) overlaps significantly in meaning with the explicit criticism "堆砌词藻" (piling up ornate words), failing to add new information. More importantly, it might have lost the rich expression of the original sentence.
• **LTP Model:** 这份报告言之无物，只是堆砌词藻。(This report is devoid of substance, merely piling up words.)
Reasoning: Correctly identifies the core implicit flaw to be addressed is "内容空洞" (vacuous content), rather than the already explicit critique of style ("堆砌词藻"). Selects "言之无物" (to be devoid of substance) which directly and accurately targets this core issue without redundancy, complementing rather than repeating the existing criticism.

Table 10: Success Cases: Comparative examples demonstrating the improved capabilities of the LTP-finetuned model over the Base model across different aspects of the CIP task.

Optimizable Case Analysis

A. Cultural Nuance & Metaphor

• **Original:** 这位老将军退休回归生活之后，生活十分朴素。 (After retiring and returning to civilian life, this old general lived very simply.)
• **LTP Model:** 这位老将军布衣蔬食，生活十分朴素。(This old general wore cloth and ate simple food, living very simply.)
• **Error Analysis:** The idiom "布衣蔬食" literally describes a simple lifestyle of wearing coarse clothes and eating plain food. However, its cultural connotation is strongly associated with the deliberate, virtuous austerity of scholars or commoners. Applying it to a retired military general ignores the specific cultural metaphor encapsulated in the more appropriate idiom "解甲归田". While "布衣蔬食" describes how one lives, it misses the crucial nuance of identity transition from a warrior ("甲" armor) to a farmer ("田" field), which is central to the original context.

B. Complex Semantic Compositionality

• **Original:** 这个新方案巧妙地绕开了所有技术难点，令人赞叹。 (This new scheme cleverly circumvented all technical difficulties, which was admirable.)
• **LTP Model:** 这个新方案另辟蹊径，令人赞叹。(This new scheme blazed a new trail, which was admirable.)
• **Error Analysis:** The idiom "另辟蹊径" (to blaze a new trail) captures the novelty" and cleverness but misses the core action of "circumventing difficulties". The context demands an idiom that holistically represents the clever circumvention of difficulties by the scheme. The model's output satisfies only a subset. A more precise idiom could be "化繁为简" (simplify the complex).

C. Discourse-Level Coherence

• **Original Context:** (Previous sentences describe a team persevering through multiple failures.) 虽然失败了很多次，但他们从不灰心，这次又开始了新的尝试。 (Although they failed many times, they never lost heart, and this time they started a new attempt.)
• **LTP Model:** 虽然失败了很多次，但他们从不灰心，这次又卷土重来。(Although they failed many times, they never lost heart, this time they stage a comeback.)
• **Error Analysis:** In isolation, "卷土重来" means to return after a defeat and can be used neutrally. However, it often carries a **negative or adversarial connotation** (e.g., a defeated enemy regrouping). This clashes with the positive, persevering tone established for the team in the preceding discourse. The substitution disrupts the narrative's emotional consistency. A more appropriate choice would be "重整旗鼓" (reorganize and rally, more neutral) or "再接再厉" (make persistent efforts).

D. Rare Idioms & Data Sparsity

• **Original:** 他们夫妻二人早年间生活极其贫困，只能共用一件简陋的蓑衣御寒，却始终相互扶持。(In their early years, the couple lived in extreme poverty, sharing only a crude straw raincoat for warmth, yet they always supported each other.)
• **LTP Model:** 他们夫妻二人早年间生活一贫如洗，却始终相互扶持。(In their early years, the couple lived in utter poverty, yet they always supported each other.)
• **Error Analysis:** The model selects the common idiom "一贫如洗" (penniless), which only captures the generic state of poverty, entirely omitting the iconic and narrative-rich detail of "共用一件简陋的蓑衣御寒" (sharing a crude straw coat for warmth). This detail is precisely and uniquely described by the rare idiom "牛衣对泣" (literally, to face each other and weep in an ox-hide garment), which originates from a specific historical anecdote about a poor couple and has become the sole, fixed expression for describing a couple enduring harsh poverty together.

Table 11: Optimizable Cases: Analysis of typical failure modes of the LTP-finetuned model, highlighting remaining challenges and pointing to concrete future research directions.