PhonoThink: Improving Large Language Models' Reasoning on Chinese Phonological Ambiguities

Jianfei Ma* **and Zhaoxin Feng*** **and Emmanuele Chersoni**
and **Huacheng Song** and **Ziqi Zhang**
Department of Language Science and Technology, The Hong Kong Polytechnic University
{jian-fei.ma,zhaoxinbetty.feng,huacheng.song}@connect.polyu.hk,
emmanuele.chersoni@polyu.edu.hk,
blameredens.zhang@connect.polyu.hk

Abstract

Effectively resolving phonological ambiguities is crucial for robust natural language processing, as these ambiguities are pervasive in tasks ranging from speech-to-text, spelling correction, to offensive language detection. However, current Large Language Models (LLMs) frequently struggle to resolve such ambiguities.

To address this challenge, we present a framework to enhances LLMs' phonological capability through a multiple-stage training approach. Our method begins with supervised fine-tuning on well-constructed datasets, including three subtask datasets designed to enhance the model's foundational phonological knowledge, along with a synthetic dataset of step-by-step reasoning chains. Following this, we apply reinforcement learning to incentivize and stabilize its reasoning.

Results show that our framework enables the base model to achieve relatively comparable performance to a much larger model. Our ablation studies reveal that subtask datasets and the synthetic dataset can simultaneously impact as complementary modular enhancers to strengthen LLMs' integrated application[1].

1 Introduction

Phonological information, a key component of language, interacts dynamically with semantic meaning to form a coupled system (de Saussure and Baskin, 2011). It plays an empirically critical role in diverse downstream tasks, such as spelling correction (Liu et al., 2024; Li et al., 2024b; Liu et al., 2025), offensive language detection (Xiao et al., 2024), and pun generation (Jaech et al., 2016; Xu et al., 2024). This is also particularly true in Chinese, which features many-to-one relationships between characters and pronunciation and

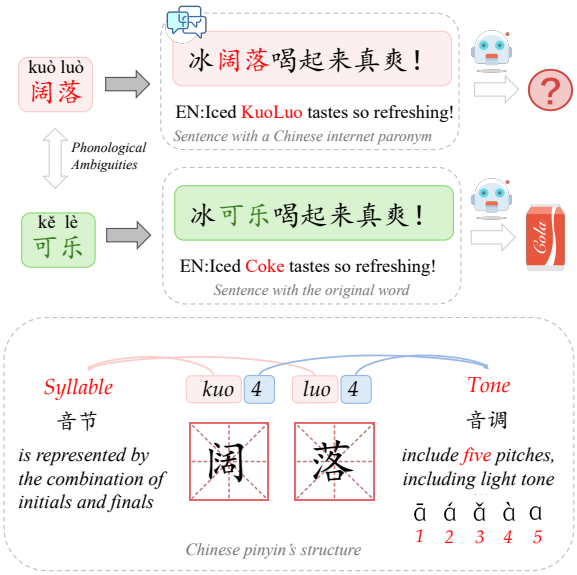

Figure 1: Introduction of the structure of Chinese *pinyin* with an example of phonological ambiguity.

all the related tasks are fundamentally in the context of phonological similarity between Chinese words — like how the internet slang replaces "可乐" (*Coke*) with a paronym or near-homophone[2] "阔落" (*KuoLuo*, just referring to Coke but with no inherent meaning in Chinese) to convey relaxed, humorous and coquettish tone. This showcases how phonetic equivalence or similarity drive linguistic creativity and ambiguity, as shown in Figure 1.

Although recent Large Language Models (LLMs) have demonstrated impressive performance, frequently eclipsing human abilities across various tasks and domains (Brown et al., 2020; Chen et al., 2021a; Lewkowycz et al., 2022; Luo et al., 2025), their capabilities still face significant challenges when facing Chinese phonologi-

* represents these authors contributed equally to this work.
[1]Our dataset and code is available at the link.

[2]According to Unseth and Unseth (1991) and Valera and Ruz (2021), paronyms and homophones are words carrying different meaning but pronounced or written similarly (paronyms) or identically (homophones). They can often lead to confusion, like "complement" (to complete) and "compliment" (a praise) in English.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 19019–19034
November 4-9, 2025

Figure 2: Comparison between reasoning chains of Qwen2.5-7B and those of the model reinforced with our PhonoThink strategy (since the model generates responses exclusively in Chinese, we include their English translations here to facilitate reader comprehension, see Chinese version in Appendix A).

cal ambiguities where embedded semantic under homophones lead to comprehension failures (Xiao et al., 2024). The root cause of this performance gap lies in LLMs' inadequate awareness of Chinese phonological variants who can potentially disrupt semantic grounding due to homophones or paronyms obscure intended meanings. Such deficiencies not only lead to literal misinterpretations, but also hinder higher-level reasoning where phonological awareness is crucial (e.g., resolving puns or detecting offensive homophone substitutions).

It has been suggested that native Chinese speakers leverage their perceptual systems to retrieve original words from homophonic variants through phonological similarity-based reasoning and contextual information understanding (Samuel, 1981; Davis et al., 2005; Banfi and Arcodia, 2013; Mehta and Luck, 2020). Building upon this human cognitive paradigm, our study proposes a pipeline to equip LLMs with analogous reasoning capabilities.

By analyzing existing LLMs' performance in identifying and restoring homophones/paronyms (hereafter termed **target words**) in sentences (see Figure 2), we identify three cruxes (1) the failure to precisely identify the target words (e.g., incorrectly capturing “冰阔落” (*iced KuoLuo*) with its modifier, instead of just “阔落” (*KuoLuo*)); (2) the inadequate conversion of Chinese words to *pinyin*[3] representations; and (3) the inability to generate sufficiently diverse phonetically similar candidates (often exhibiting repetitive outputs). The detailed performances of these issues are demonstrated in the Appendix E. These deficiencies collectively impair LLMs' capacity to accurately restore target words to their corresponding original words.

To address current LLMs' shortcomings in *pinyin* conversion, phonetically similar words as-

[3] *Pinyin* is a Latin-based phonetic notation system for Chinese, representing character pronunciation through syllables and tones, as shown in Figure 1.

sociation, and strengthen LLMs' performance in context-based masked word prediction, we first constructed three subtask datasets. To enhance the models' capability in processing sentences with phonological ambiguities, we further developed a four-stage reasoning dataset by utilizing a synthetic dataset. In our implementation, we employed the Qwen2.5-72B model (Qwen et al., 2025) as the teacher model to generate reasoning content step by step, which was then integrated into complete reasoning chains as training seed data.

After strategically integrating the reasoning dataset with three subtask datasets, we then conducted a Supervised Finetuning (SFT). Also, we employed Group Relative Policy Optimization (GRPO) (Shao et al., 2024) for Reinforcement Learning (RL) by combining synthetic data with 20% stratified samples from each authentic dataset, while implementing a compound reward system comprising strict XML formatting rewards and four step-specific correctness rewards.

Experimental results demonstrate that the base model (Qwen2.5-7B-Instruct) exhibited a nearly 50% enhancement in accuracy when evaluated on our synthetically curated dataset and significantly outperforms a R1-distilled reasoning model of the same size. This validates the efficacy of our training pipeline in enhancing LLMs' reasoning capabilities for Chinese phonological ambiguities.

2 Related Work

Phonological ambiguities involve words sounding the same (homophones), or similarly (paronyms), but having different meanings (Unseth and Unseth, 1991; Valera and Ruz, 2021). Humans can parse homophones or paronyms to understand the originally underlying meanings through a psychological effect named perceptual compensation (Samuel, 1981; Davis et al., 2005; Banfi and Arcodia, 2013; Mehta and Luck, 2020). However, LLMs often struggle to comprehend sentences where words have been substituted with their homophones or paronyms, demonstrating their limited robustness in phonological perturbations (Xiao et al., 2024).

Current research on the ability of LLMs to comprehend phonological information remains limited. In the integrated evaluation of LLMs' phonological ability, a previous study conducted explorations on syllable counting, rhyme generation, and grapheme-to-phoneme, finding that LLMs' phonological capabilities are relatively low (Suvarna et al., 2024). Focusing on single tasks, the involvement of the phonetic spelling system in phonetic symbol correction (Qharabagh et al., 2024) and character correction (Li et al., 2024c; Tang et al., 2024) can considerably enhance the performance. Based on the above findings, we designed a strategy to optimize the knowledge of the phonetic spelling system (*pinyin*) in reasoning tasks related to phonological ambiguity.

To evaluate our method's effectiveness in improving LLMs' handling of Chinese phonological ambiguities, we employ three benchmark tasks: (1) Chinese Typing Correction (CTC) (Zhu et al., 2022; Li et al., 2024c,a), (2) Automatic Speech Recognition (ASR) Transcript Correction (Liu et al., 2025; Wei et al., 2024), and (3) Chinese Internet Homophone and Paronym Restoration.

3 Methodology

3.1 Problem Definition

Let $D = \{(X, Y)\}$ denote a dataset where each consists of a sentence X with a homophone or paronym and the corresponding original word Y.

The task of LLM is to analyze X through a four-step reasoning process, where s_i represents the reasoning step i. Specifically, s_1 involves detecting the target word (homophone or paronym) w in X; s_2 converts w into its *pinyin* representation p; s_3 generates a list of candidate words $C = \{c_1, c_2, \ldots, c_m\}$ that are contextually appropriate based on the context of X and phonetically similar to w; and s_4 selects the word that is most likely to be the original word $\hat{Y}$ from C. Formally, the output of the critic can be represented as:

$$(s_1, s_2, s_3, s_4) \sim \pi_\theta(X), \tag{1}$$

The goal is to ensure $\hat{Y} = Y$, meaning that the LLM correctly restores the target word. In this study, we propose a multi-stage methodology to achieve the objective. First, we construct three subtask-specific datasets aimed at enhancing the LLM's capabilities in Chinese *pinyin* conversion, homophones and paronym association, and context-based masked word prediction. Concurrently, we create a synthetic dataset encompassing comprehensive reasoning processes. These four datasets are then combined for SFT. Subsequently, we employ RL to provide stepwise rewards for the LLM's reasoning process. The framework of our proposed approach is illustrated in Figure 3.

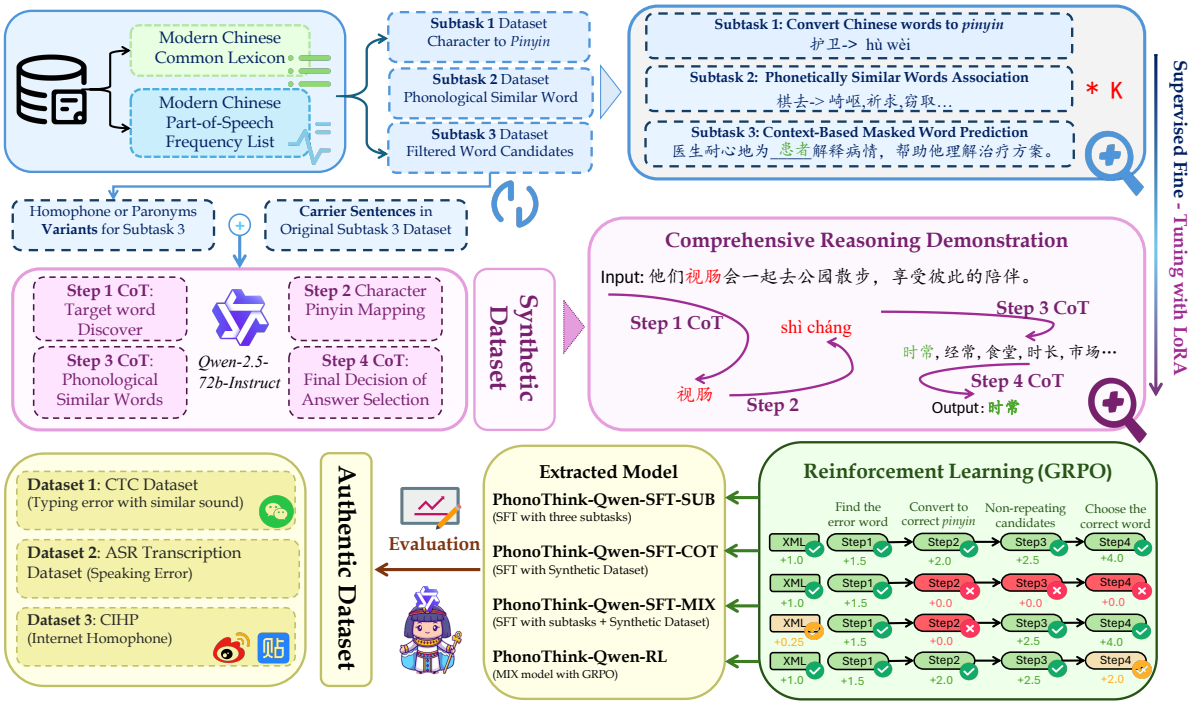

Figure 3: Pipeline of our proposed PhonoThink strategy. The same color indicates same corresponding procedure. The blue stage represents the construction of the Subtasks Dataset. The pink or purple stages denote the building of synthetic datasets for SFT and GRPO, respectively. The green block illustrates the reward logic of GRPO. The yellow stage shows the final trained models and how they are incorporated into the evaluation process.

3.2 Dataset Construction

Subtask Datasets As illustrated in the blue stage in Figure 3, we developed three subtasks datasets.

- **Subtask 1 dataset**, derived from the *Modern Chinese Common Lexicon* and processed via Python's *pypinyin* library[4], trains models to learn to convert Chinese words to *pinyin*.
- **Subtask 2 dataset**, constructed from 5,000 disyllabic words in *Modern Chinese Common Lexicon*, enables phonetic similarity ranking through *fuzzywuzzy* scoring[5] and IPA-based refinement. This dataset is designed to equip LLMs with the ability to recognize phonetic similarity hierarchically, covering across homophones and paronyms.
- **Subtask 3 dataset**, filtered from the *Modern Chinese POS Frequency List*[6] to 5,596 contextually relevant words, supports masked word prediction; its sentences were synthetically generated by Qwen2.5-72B-Instruct and manually validated.

Specifically, an example from the subtasks dataset is provided in the right of the blue-shaded section in Figure 3. Besides, On the far right of the figure, the top-down arrow illustrates the multi-task SFT process, which is performed on a combined dataset of subtasks and synthetic data. For detailed construction procedures, see Appendix G.

Synthetic Datasets The synthetic dataset was specifically constructed for SFT and RL training purposes. The existing related Chinese datasets predominantly focus on homophones with identical syllable structures, while lacking coverage of paronyms (similar but non-identical syllables). So we systematically generated synthetic data by modifying the sentences from subtask 3, replacing original words with homophones or paronyms across varying degrees of phonetic similarity.

This process involved four steps: 1) computing *fuzzywuzzy* similarity scores for all lexicon entries, 2) applying a 5:2:1:1:1 ratio to categorize words into tiers before IPA-based re-ranking, 3) substituting optimal matches with artificial phonetic equiv-

[4]*Pypinyin* is a Python library to convert Chinese characters to *pinyin* transcriptions with tone marks, available at link.

[5]*Fuzzywuzzy* is a Python library for efficient string matching and similarity calculation using Levenshtein distance, commonly applied in text processing and phonetic comparison tasks, available at link.

[6]This corpus was sourced from Professor Xing Hongbing at Beijing Language and Culture University, available at link.

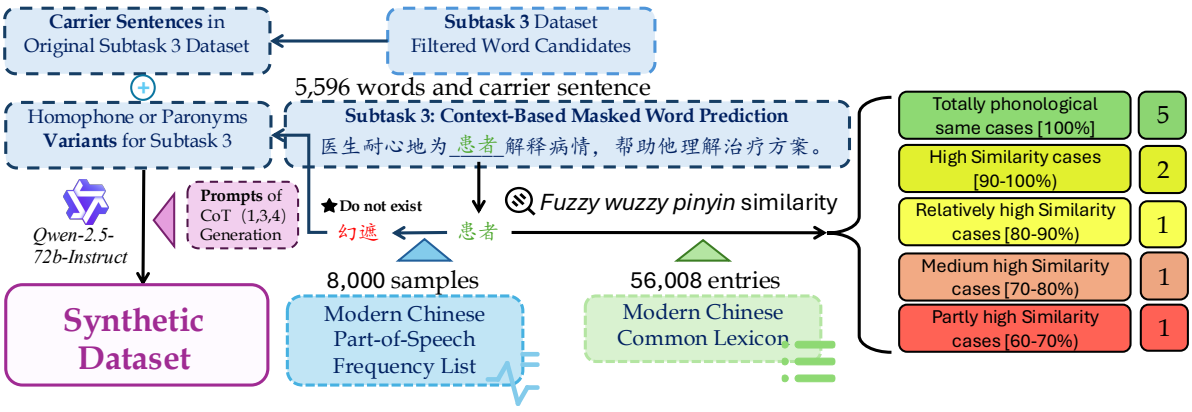

Figure 4: Pipeline of Synthetic Dataset Construction and Part of Step 3 SFT Dataset. This figure presents a micro-level view of the construction process for the most essential synthetic dataset. The colors correspond to those in Figure 3, with the same color indicating the same procedure. However, the numbers on the right represent sample counts rather than procedural steps.

alents, and 4) generating the standardized triplet structure (target word, original word, contextual sentence). For detailed process, please refer to Figure 4 and the textual description in Appendix G.

Authentic Datasets To evaluate LLMs' performance on authentic scenarios, we constructed three real-world datasets: Chinese Typing Correction, Automatic Speech Recognition Transcript Correction, and Chinese Internet Homophones and Paronyms Restoration. The first two datasets were curated from existing resources, with the first dataset predominantly comprising perfect homophones (phonologically identical pairs), while the third dataset was manually collected from Weibo and Tieba[7] (contains both homophones and paronyms).

The CSCD-NS dataset (Hu et al., 2024) comprises 40,000 real-world Chinese typing errors from Weibo, annotated at multiple linguistic levels (character/phrase) for comprehensive typing correction evaluation. We randomly sampled 1,000 disyllabic errors, validated by three native speakers, and retained 836 contextually appropriate cases where the original word maintained semantic coherence with the surrounding text. This forms our first authentic dataset.

The AISHELL-3 corpus (Shi et al., 2021), comprising 85 hours of annotated Mandarin speech from 218 native speakers, was processed using Kaldi-based ASR (Chen et al., 2021b) and converted to simplified Chinese. Following protocols established for CSCD-NS, we randomly sampled 1,000 entries and extracted disyllabic words while filtering context-dependent errors through tripartite native-speaker verification, yielding 387 validated error instances for the second authentic dataset.

The third dataset, CIHP (Ma et al., 2025), consists of target homophones and paronyms extracted from Weibo and Tieba. To create this dataset, a systematic analysis of randomly sampled user-generated content spanning from 2010 to early 2025 (pre-March cutoff) was conducted. From this content, three native Chinese annotators collaboratively identified 352 high-frequency, representative pairs of target and original words. Each pair was included in the final dataset only after reaching a consensus among all three annotators, along with its corresponding *pinyin* representation.

3.3 Teaching LLMs to Reason

To enhance LLMs' reasoning capabilities for dealing with phonological ambiguities, we constructed a four-stage reasoning chain SFT dataset based on synthetic training data. Three simple subtask datasets for learning basic knowledge to better match each step answers' form, and one reasoning chain dataset for teaching step-by-step phonological ambiguity resolution. For the reasoning chain data, we employed Qwen2.5-72B-Instruct to generate reasoning processes for steps 1, 3, and 4, then consolidated them into reasoning chains[8]. After-

[7] Weibo, managed by Sina company, is a popular Chinese microblogging platform similar to Twitter and Tieba, hosted by Baidu, is a large online community forum where users can engage in topic-based discussions, akin to Reddit.

[8] In the paper, we state that Step 2 does not require a reasoning chain because *pinyin* conversion is a deterministic, rule-based process that essentially involves only character-to-*pinyin* mapping, with no need for inference.

wards, through SFT, the base model learned both the structural format and logical reasoning process. The detailed procedure is as follows:

Manual Reasoning Exemplars Three native Chinese speakers collaboratively created detailed reasoning exemplars for steps 1, 3, and 4 steps, which were then integrated into prompt templates to guide the teacher model's reasoning generation.

Comprehensive Reasoning Chains For the synthetic dataset $\mathcal{D} = \{(X, w, Y, \mathcal{L})\}$, where each instance contains a sentence X with target word w, its corresponding original word Y and data usage condition $\mathcal{L}$[9]. Using Qwen2.5-72B-Instruct (θ^*), we generate targeted reasoning for each critical step through isolated prompting, respectively:

$$r_i = \pi_{\theta^*}(\cdot|X, w, \mathcal{L}), \quad i \in \{1, 3, 4\} \tag{2}$$

where r_i denotes the generated reasoning for step s_i. The complete reasoning chain is then synthesized by combining these step-specific outputs with the ground-truth *pinyin* conversion p:

$$\mathcal{R}_{\text{complete}} = (r_1, p, r_3, r_4) \tag{3}$$

Here, Ψ represents the integration function to generate complete reasoning chain. All chains undergo manual validation to ensure reasoning fidelity, with annotators correcting instances where generated outputs don't match intended corrections. Through this pipeline, we obtained approximately 2.9K reasoning chain data.

Supervised Fine-Tuning We integrated reasoning chains with the three subtask datasets at varying ratios to construct a composite dataset D_{SFT} to be subsequently utilized to SFT for enhancing the model ' s reasoning capabilities in resolving phonological ambiguities. The initial reasoning model, PhonoThink-Qwen-SFT-MIX θ_{SFT}, was obtained:

$$\theta_{SFT} = \arg\min_{\theta} \mathbb{E}_{(X,w,\mathcal{R})\sim\mathcal{D}_{SFT}} \tag{4}$$

[9]The variable $\mathcal{L}$ represents constraint conditions applied during reasoning chain generation at different steps. Specifically, in Step 1, $\mathcal{L}$ consists of only the sentence containing the homophone, with other dataset elements excluded from the input prompt. In Step 3, $\mathcal{L}$ includes the sentence with the homophone, the homophone itself, and its *pinyin* representation. In Step 4, $\mathcal{L}$ comprises the sentence with the homophone, the homophone, and the original word candidates generated from Step 3.

3.4 Incentivizing LLMs to Reason Better

Upon obtaining the initial reasoning model, we employ continued incentivization through GRPO (Shao et al., 2024) to fully elicit its potential. The training incorporates approximately 1K samples from the synthetic dataset (excluding those used for SFT) combined with 20% stratified samples from each of the three authentic datasets[10], creating a composite training set for GRPO optimization. In terms of reward function design, we developed a total of five distinct reward functions: (1) a strict XML formatting reward function, accompanied by (2) four step-specific correctness reward functions corresponding to each reasoning step. The illustration is visually in Figure 3's reinforcement learning stage.

4 Experiments and Analysis

4.1 Experimental Settings

Base Model To evaluate the proposed pipeline for enhancing reasoning, we adopted Qwen2.5-7B-Instruct (Qwen et al., 2025) as the base model.

Tasks This pipeline passes two training stages: 1) Multi-task SFT on three subtasks and detailed CoT reasoning chain to yield PhonoThink-Qwen-SFT; and 2) GRPO optimization to incentivize reasoning ability to give PhonoThink-Qwen-RL.

Baselines To compare our model in a multi-dimensional perspective, we choose two large-scale models, Qwen2.5-72B and GPT-4o, which are likely to represent an upper bound for our task performance. Plus, we chose a similarly scaled distill model, DeepSeek-Distill-Qwen-7B, to compare out strategy to what is achievable with a model with default reasoning ability.

Configurations During SFT, we set the learning rate to 3e-5, train for 2 epochs with a batch size of 4, using LoRA with rank 16 and alpha equal to 16. In GRPO, we set the learning rate as 3e-5 to train the model in 1 epoch, using Lora with rank 16 and alpha 16. For evaluation, we generate outputs with temperature and top-p equal to 0.5, using max token lengths of 2048, and are directly measured by accuracy. All the experiments are run on one NVIDIA A100-40GB GPU.

[10]Since we utilized 20% of the three authentic datasets for reinforcement learning, the testing on authentic datasets was conducted using the remaining 80% of the original datasets.

Model Name	Synthetic Data	CTC	ASR	CIHP
Qwen2.5-7B-Instruct	16.98%	3.89%	3.24%	20.57%
Deepseek-Distill-Qwen-7B	4.65%	1.95%	0.65%	6.03%
Qwen2.5-72B-Instruct	71.86%	13.17%	28.16%	46.45%
GPT-4o	78.60%	48.35%	36.57%	59.93%
PhonoThink-Qwen-SFT-SUB	23.78%	2.40%	0.52%	6.03%
PhonoThink-Qwen-SFT-COT	31.47%	8.98%	1.29%	20.45%
PhonoThink-Qwen-SFT-MIX	60.84%	13.47%	8.74%	27.66%
PhonoThink-Qwen-RL	66.21%	18.25%	16.17%	35.79%

Table 1: This table compares the performance of the base model (Qwen2.5-7B-Instruct), baseline models, and our proposed models across four test datasets. PhonoThink-Qwen-SFT-SUB is the base model SFT-finetuned on three subtasks, PhonoThink-Qwen-SFT-COT is SFT-finetuned only on the CoT dataset, PhonoThink-Qwen-SFT-MIX uses mixed-dataset SFT, and PhonoThink-Qwen-RL further optimises PhonoThink-Qwen-MIX via RL. The pink, yellow, and blue background colors correspond to the base, baseline, and our proposed models, respectively.

4.2 Experimental Results

Table 1 shows the performance of different models in this experiment on synthetic data and three authentic datasets.

Base Model Qwen2.5-7B-Instruct exhibited substantially subpar performance on all four evaluation datasets, with notably dismal accuracy rates falling below 5% specifically on the CTC and ASR benchmarks. Interestingly, the performance of DeepSeek-Distill-Qwen-7B is even lower, suggesting that default reasoning is not sufficient for our tasks.

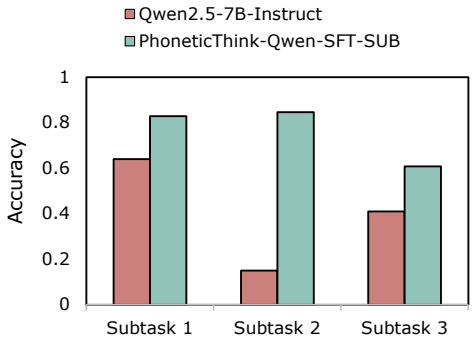

Figure 5: Comparison of performance of the base model and models after SFT using three subtask datasets.

4.2.1 Results of Supervised Finetuning

SFT with Subtask Datasets To validate the effectiveness of our subtask datasets in enhancing the base model's three critical sub-skills (*pinyin*-character mapping, association of phonetically similar words, and context-based masked word prediction, which are essential for integrated tasks), we first conducted SFT on the base model only using three subtask datasets. To ensure balanced sample sizes across datasets, given that the training set of subtask 2 (with the smallest size) contains 3,615 instances, we randomly sampled 3,615 data points from each of the other two subtask training sets. These were subsequently constructed into a consolidated SFT dataset comprising all three subtasks.

As shown in Figure 5, SFT models on these three subtasks yield consistent improvements, suggesting the potential accessibility for further challenging reasoning. For subtask 2, we can closely observe significant advancement in several aspects of the model's responses, including reduced lexical repetition and expanded vocabulary selection.

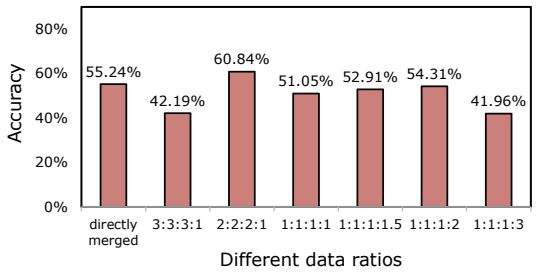

Figure 6: Model performance on the synthetic test set after SFT with datasets of varying composition ratios.

SFT with Mixed Dataset We combined the reasoning chain dataset with the merged three-subtask dataset in the last step to create a mixed dataset for SFT on the base model. After testing different data ratios (Figure 6), we find that a balanced 2:2:2:1 ratio yields optimal performance with highest accuracy, significantly outperforming both the directly merged approach[11] and other imbalanced

[11]The "directly merged approach" refers to the straight-

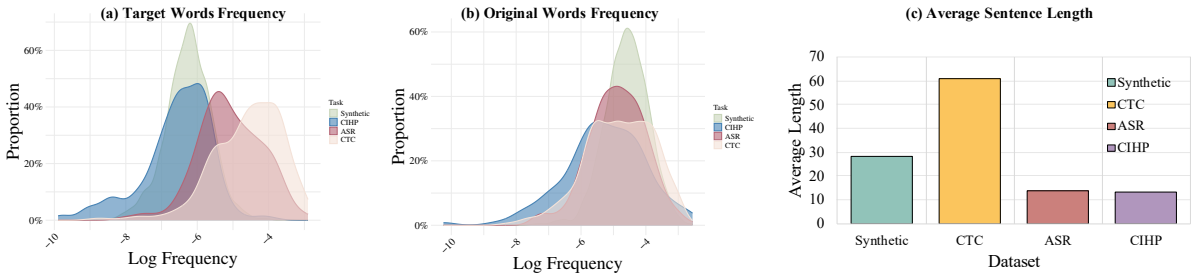

Figure 7: Log frequency distributions of target and original words and average sentence lengths of different datasets.

configurations. Results also suggest that including either too much or too little of the reasoning chain dataset in the training mixture negatively impacts the model's performance on this task.

With the mixed dataset with a 2:2:2:1 ratio, our final PhonoThink-Qwen-SFT-MIX model is established. Then, assessed on three authentic datasets, the performance is detailed in Table 1. A comparison with the base model reveals that multi-task learning, combining three subtasks and CoT processes, significantly enhances the LLM's ability by jointly leveraging phonetic knowledge and reasoning patterns. Moreover, compared to only using the CoT dataset for SFT (PhonoThink-Qwen-SFT-COT), the mixed-dataset SFT approach demonstrates significantly superior performance across all four datasets. This suggests that a combined SFT strategy incorporating both subtask and integrated reasoning training is more beneficial for complex reasoning tasks.

This enhancement observed after introducing the CoT dataset likely arises from the model's ability to emulate human-provided reasoning patterns. The CoT framework, which combines step-by-step reasoning with phonetic information, facilitates integrated learning of both knowledge types, resulting in superior performance compared to models trained exclusively on basic phonetic knowledge.

However, the performance gap between synthetic and authentic datasets remains significant due to distributional discrepancies in (1) sentence length and (2) linguistic properties of target words, including their target/original words' frequency and POS distributions. This explains why the substantial improvements observed on synthetic data do not fully generalize to authentic datasets. For the distribution of target/original words' frequency and sentence length, refer to Figure 7; for POS distribution, see Appendix B.

forward combination of the previously constructed subtask composite dataset with the reasoning training dataset at a fixed ratio of 3615:2615:3615:2994 without further balancing.

4.2.2 Results of Reinforcement Learning

GRPO training significantly stabilizes the model's reasoning process by providing step-wise rewards, enabling finer-grained optimization. As evidenced in Table 1, PhonoThink-Qwen-RL demonstrates superior performance compared to PhonoThink-Qwen-SFT-MIX across four datasets.

Notably, PhonoThink-Qwen-RL substantially outperforms the base model and DeepSeek-R1-Distill-Qwen-7B (a reasoning-capable model of comparable parameter size), highlighting that our multi-stage approach is empirically effective. Remarkably, our method enables the base model (Qwen2.5-7B-Instruct) to approach the performance of a much larger-scale model (Qwen2.5-72B-Instruct), effectively overcoming the parameter efficiency barrier for the restoration of Chinese phonological ambiguities. At last, particially due to that we only incorporated a small amount of real-world data during reinforcement learning, PhonoThink-Qwen-RL demonstrates marginal improvements across all three authentic datasets, and its overall performance remains relatively low.

5 Conclusion

This study proposes a multi-stage pipeline that activates language models to leverage both phonetic and semantic information. Using Qwen2-7B-Instruct as our base model, we demonstrate the effectiveness of this approach through our synthetic dataset and three authentic datasets.

Our evaluation shows SFT with subtask datasets enhances the model's ability to use phonological information for basic phonetic tasks. By integrating SFT with CoT and subtasks datasets, enabling explicit learning of reasoning patterns, the model achieves performance comparable to larger LLMs with the same architecture. GRPO further enhances

output quality through fine-grained reward during stepwise reasoning, achieving results comparable to larger models with CoT prompting.

However, due to distinction in training data scale and authentic task features, the performance increase on the authentic tasks is more limited, which will be a direction of work for future studies.

Limitations

Our proposed workflow validated only on Chinese; cross-lingual adaptability remains unverified. Besides, the scope of our validation does not extend to other phonetically-intensive tasks (e.g., metered poetry generation and prosodic pattern modeling) that require deeper phonological reasoning.

Furthermore, the autoregressive LLMs are constrained by semantic conflicts caused by target words, which disrupt the mapping from characters to words (Li et al., 2024a), amplifying segmentation errors and leading to error accumulation, ultimately resulting in entirely flawed reasoning. Our method can not solve this problem. Additionally, the capacity ceiling of Qwen2.5-72B-Instruct restricts the quality of reasoning chains. Due to computational limitations, we were unable to test more powerful models. Larger architectures with multilingual capabilities could potentially enhance performance across different languages and tasks.

The final limitation lies in our synthetic dataset, constructed using *fuzzywuzzy* and IPA-based phonetic distance metrics, which may deviate from real-world error patterns and overlook phonological phenomena such as elision and liaison. Moreover, with just over 5,000 data entries, the dataset is relatively small compared to typical NLP tasks. This limited size prevents the dataset from encompassing errors of various types, consequently creating a discrepancy between synthetic data and authentic data.

Ethics Statement

We do not foresee any immediate negative ethical consequences of our research.

References

Emanuele Banfi and Giorgio Francesco Arcodia. 2013. On line proceedings of the sixth mediterranean morphology meeting the shng/sheng complex words in chinese between morphology and semantics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021a. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Yi-Chang Chen, Chun-Yen Cheng, Chien-An Chen, Ming-Chieh Sung, and Yi-Ren Yeh. 2021b. Integrated semantic and phonetic post-correction for Chinese speech recognition. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 95–102, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Matthew H Davis, Ingrid S Johnsrude, Alexis Hervais-Adelman, Karen Taylor, and Carolyn McGettigan. 2005. Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *J. Exp. Psychol. Gen.*, 134(2):222–241.

Ferdinand de Saussure and Wade Baskin. 2011. *Course in General Linguistics: Translated by Wade Baskin. Edited by Perry Meisel and Haun Saussy*. Columbia University Press.

Yong Hu, Fandong Meng, and Jie Zhou. 2024. CSCD-NS: a Chinese spelling check dataset for native speakers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–159, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Jaech, Rik Koncel-Kedziorski, and Mari Ostendorf. 2016. Phonological pun-derstanding. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 654–663, San Diego, California. Association for Computational Linguistics.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

Kunting Li, Yong Hu, Liang He, Fandong Meng, and Jie Zhou. 2024a. C-LLM: Learn to check Chinese spelling errors character by character. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5944–5957, Miami, Florida, USA. Association for Computational Linguistics.

Yuang Li, Xiaosong Qiao, Xiaofeng Zhao, Huan Zhao, Wei Tang, Min Zhang, and Hao Yang. 2024b. Large language model should understand pinyin for chinese asr error correction. *arXiv preprint arXiv:2409.13262*.

Yuang Li, Xiaosong Qiao, Xiaofeng Zhao, Huan Zhao, Wei Tang, Min Zhang, and Hao Yang. 2024c. Large language model should understand pinyin for chinese asr error correction. *Preprint*, arXiv:2409.13262.

Changchun Liu, Kai Zhang, Junzhe Jiang, Zixiao Kong, Qi Liu, and Enhong Chen. 2025. Chinese spelling correction: A comprehensive survey of progress, challenges, and opportunities. *Preprint*, arXiv:2502.11508.

Changchun Liu, Kai Zhang, Junzhe Jiang, Zirui Liu, Hanqing Tao, Min Gao, and Enhong Chen. 2024. ARM: An alignment-and-replacement module for Chinese spelling check based on LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10156–10168, Miami, Florida, USA. Association for Computational Linguistics.

Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O Cohen, Valentina Borghesani, Anton Pashkov, and 1 others. 2025. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, 9(2):305–315.

Jianfei Ma, Zhaoxin Feng, Huacheng Song, Emmanuele Chersoni, and Zheng Chen. 2025. Reasoning or memorization? investigating LLMs' capability in restoring Chinese Internet homophones. In *Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM)*, pages 120–139, Vienna, Austria. Association for Computational Linguistics.

Anita Mehta and Jean-Marc Luck. 2020. Hearing and mishearings: Decrypting the spoken word. *Advances in Complex Systems*, 23(03):2050008.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Mahta Fetrat Qharabagh, Zahra Dehghanian, and Hamid R. Rabiee. 2024. Llm-powered grapheme-to-phoneme conversion: Benchmark and case study. *Preprint*, arXiv:2409.08554.

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, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

A G Samuel. 1981. Phonemic restoration: insights from a new methodology. *J. Exp. Psychol. Gen.*, 110(4):474–494.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *Preprint*, arXiv:2010.11567.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. PhonologyBench: Evaluating phonological skills of large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Zhiyuan Tang, Dong Wang, Shen Huang, and Shidong Shang. 2024. Pinyin regularization in error correction for chinese speech recognition with large language models. In *Interspeech 2024*, pages 1910–1914, ISCA. ISCA.

Carole Unseth and Peter Unseth. 1991. Analyzing ambiguity in orthographies. *Notes on Literacy*, 65:35–52.

Salvador Valera and Alba E Ruz. 2021. Conversion in english: homonymy, polysemy and paronymy. *English Language & Linguistics*, 25(1):181–204.

Victor Junqiu Wei, Weicheng Wang, Di Jiang, Yuanfeng Song, and Lu Wang. 2024. Asr-ec benchmark: Evaluating large language models on chinese asr error correction. *Preprint*, arXiv:2412.03075.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6012–6025, Miami, Florida, USA. Association for Computational Linguistics.

Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. "a good pun is its own reword": Can large language models understand puns? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11766–11782, Miami, Florida, USA. Association for Computational Linguistics.

Chenxi Zhu, Ziqiang Ying, Boyu Zhang, and Feng Mao. 2022. MDCSpell: A multi-task detector-corrector framework for Chinese spelling correction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1244–1253, Dublin, Ireland. Association for Computational Linguistics.

A Qwen-2.5-7b-Instruct and Ours

The given demonstration in the main text is used in English. The actual output of the base model and ours are given in Figure 8.

B Datasets Details

Four datasets are applied in our study for SFT, GRPO, and test. The detailed information and utilization are shown in Table 2. We split the synthetic data into a 4:1 ratio for train and test sets. We partitioned the test set in a 3:1 ratio for SFT and RL, respectively. As discussed in the main text, systematic distribution differences exist between the synthetic and authentic data. One more POS tag distribution is given in Figure 9.

C Phonetic Similarity Algorithm

Our experiment computes phonetic similarity using both *fuzzyWuzzy* and *Panphon*, where *fuzzyWuzzy* employs Levenshtein Distance to transform similarity scores into the number of edits needed for string matching, with a threshold set at 60 (below 60 considered phonetically dissimilar and above 60 as target words). The pseudocode in Table 4. Since *fuzzyWuzzy* doesn't consider articulatory features, we further applied Panphon, demonstrating in Table 5. This method converts *pinyin* to IPA using the Dragonmapper package, then computes multiple distance metrics. Since different articulatory features contribute unevenly to phonetic perception, we adopt the weighted feature edit distance to account for these variations. However, due to the computational overhead of large-scale IPA conversion and Panphon's matrix operations on CPU, we ultimately selected *fuzzyWuzzy* despite its lack of phonological evidence, supplementing this with manual validation where only terms unanimously identified as targets by two native Chinese speakers were retained.

D Authentic Datasets Annotation

- Please check the given word pair below and determine whether they can be seen as homophones or paronyms. For instance, words like "压力" (yā lì) and "鸭梨" (yā lí), which share identical *pinyin* spellings but differ in tone, can be classified as valid target words in the given context based on sound(labeled as 1). Similarly, pairs such as "什么" (shén me) and "神马" (shén mǎ), though differing in both *pinyin* spelling and tone, are considered valid if their pronunciation in standard Mandarin is sufficiently similar to allow accurate contextual reconstruction based on phonetic cues alone. Conversely, word pairs like "压力" and "森马" (sēn mǎ), which differ in both *pinyin* spelling and tone, or "压力" and "森骂" (sēn mà), which share the same tone but differ in *pinyin* spelling, are deemed invalid (labeled as 0) if the word follows phonetic cues to restore, resulting in the semantic conflict.

- In general, when an erroneous word can be successfully restored to its original form using phonological information within an appropriate context, it should be annotated as '1'. Conversely, if the phonological restoration results in semantic inconsistency within the given context, the annotation should be '0'.

In the following, independent annotation by three native Chinese speakers who were blinded to each other's selections, with only those words unanimously identified as valid homophones or paronyms by all annotators being ultimately designated as target words. Annotators received a compensation of $5 USD for every 100 data entries labeled. The inner-annotation agreement(IAA) and consistency (with Fleiss's kappa as the metric) of Subtask 3 Dataset, ASR, and CTC Datasets are 96.09% (Fleiss' $\varkappa = 0.8568$), 82.5% (Fleiss' $\varkappa = 0.7557$) and 97.3% (Fleiss' $\varkappa = 0.9309$), respectively. For ASR and CTC Datasets, 387 and 163 cases with the "Inapplicable" label were removed .

E Cruxes Analysis of LLMs' Impairment of Restoration

In the Introduction, we demonstrate our observation that the impairment in the restoration procedures for LLMs lies in three key issues: (1) failure to identify the target error word; (2) insufficient ability to map words to *pinyin*; and (3) repetition of phonetically similar candidates. In what follows, we provide quantitative evidence to highlight the significance of selecting these issues as inner steps for optimization in Table 3.

F Prompts

Our study employed prompts at multiple stages. In the three subtasks SFT, we applied the prompt as shown in Figure X. Using Qwen2.5-72B-Instruct

Figure 8: Comparison between reasoning chains of Qwen2.5-7B-Instruct and the model reinforced with our PhonoThink strategy (Chinese version).

to generate CoT data step-by-step, which is shown in Figure 11. During baseline model evaluation, we found the base model's capability insufficient to produce correct outputs with CoT alone, necessitating a CoT+one-shot prompt approach, which is shown in Figure 12, to ensure proper formatting and processing. For subsequent SFT, GRPO, and testing phases, we exclusively used CoT prompts to prevent the smaller model from being adversely affected by repetitive examples during fine-tuning, which could otherwise lead to output errors.

G Detailed Settings and Construction Procedures of Subtask and Synthetic Datasets

G.1 Subtask Dataset

For subtask 1, we split the words in the *Modern Chinese Common Lexicon* into train and test sets with a 4:1 ratio, converting each word to *pinyin* by Python's *pypinyin* library. For subtask 2, we constructed a dataset of 5,000 randomly sampled disyllabic Chinese words from subtask 1's training set. Each word's *pinyin* was systematically evaluated for phonetic similarity against all entries in the *Modern Chinese Common Lexicon* using the *fuzz.ratio* algorithm from the *fuzzywuzzy* Python library. Candidate words were stratified into five similarity tiers: exact matches [100], near-identical (90,

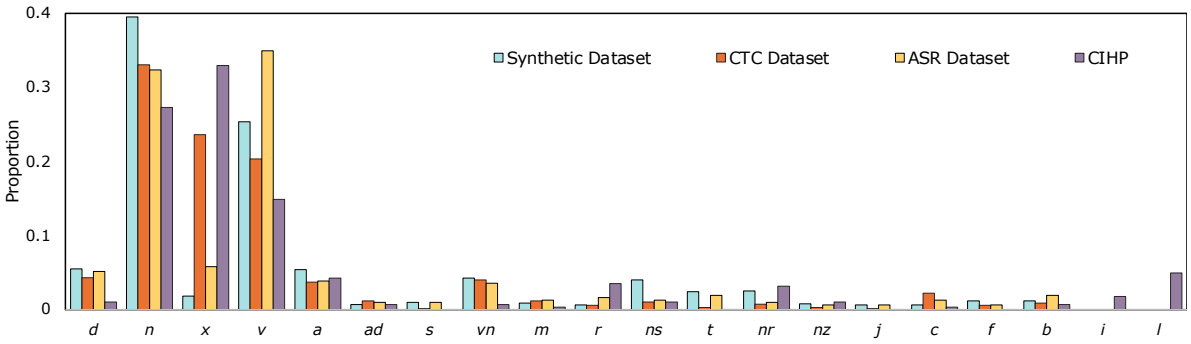

Figure 9: This figure compares POS tag distributions between the synthetic dataset and three real-world datasets, showing the percentage of each Jieba-annotated tag (y-axis) across different word categories (x-axis). While verbs and nouns dominate all datasets—and some special terms remain unidentifiable by Jieba—substantial variations in the distribution of both major and minor tags across datasets are observed. Additionally, some tag types are distributed in extremely small proportions, so we have filtered them out.

Dataset Name	**Size**	**Purpose**	**Example**
Subtask 1 Dataset	56008	SFT, test	word, *pinyin*
Subtask 2 Dataset	4519	SFT, test	original word, target word, ten words
Subtask 3 Dataset	5596	SFT, test	word, POS tag, word frequency, sentence
Synthetic Dataset	5596	SFT, GRPO, test	target word, *pinyin*, sentence, original word
CTC Dataset	836	GRPO, test	target word, *pinyin*, sentence, original word
ASR Dataset	387	GRPO, test	target word, *pinyin*, sentence, original word
CHIP Dataset	352	GRPO, test	target word, *pinyin*, sentence, original word

Table 2: This table details all the datasets in our study, specifying their sizes, purposes, and structural formats.

Crux	Qwen2.5-7B	PhonoThink-Qwen-SFT-MIX
Crux 1	with modifier: 0.1125 without modifier: 0.9375	with modifier: 0.4251 without modifier: 0.9250
Crux 2	0.6400	0.8425
Crux 3	0.6317	0.0498

Table 3: This table presents the results of key issues identified in the homophone/paronym restoration process for both the base model and the model after initial mixed SFT. The metrics for Crux 1 to 3 are accuracy, accuracy, and repetition rate, respectively. The significant performance differences between the base model and the trained model demonstrate the effectiveness of incorporating domain-specific knowledge from three mixed basic tasks into the base model.

100), high-similarity (80, 90], moderate-similarity (70, 80], and partial-similarity (60, 70]. From each tier, we initially selected 10, 4, 2, 2, and 2 candidates respectively[12], which were then rigorously evaluated through an International Phonetic Alphabet (IPA) based phonetic alignment method (Mortensen et al., 2016). The final dataset comprised the top 5, 2, 1, 1, and 1 highest-ranking candidates from each tier.

Subtask 3 is designed to enable LLMs to acquire the capability of predicting masked words through contextual inference. We utilized the annotated *Modern Chinese Part-of-Speech (POS) Frequency List*, applying sequential filters to: (1) extract the top 8,000 highest-frequency words, (2) retain only disyllabic, trisyllabic, and quadrisyllabic words

[12]The 5:2:1:1 sampling ratio was empirically determined through analysis of the Chinese Internet Homophones and Paronyms Restoration dataset in our authentic dataset, where the distribution of *fuzzywuzzy* similarity scores between target word-original word pairs across the five intervals [100], (90,100), (80,90], (70,80], and (60,70] approximately followed this proportion. The *fuzzywuzzy* algorithm was employed for initial screening due to its computational efficiency, followed by more precise IPA-based realignment.

Subtasks Prompt

Subtask 1:

【Task Instructions】

Given a Chinese word, you need to generate its standard Hanyu Pinyin (in accordance with the *Hanyu Pinyin Scheme*). Specific requirements are as follows:

1.Use spaces to separate multi-character words (e.g., "你好" → nǐ hǎo).

2. Mark tone symbols (e.g., ā á ǎ à).

3. Do not mark tones for neutral tones (e.g., "妈妈" → mā ma).

4. For polysemous characters, select the correct pronunciation based on the word's context (e.g., "银行" → yín háng, not yín xíng).

【Example】

Input word: 智能

Pinyin: zhì néng

【Input】

Input word: {item}

Pinyin:

Subtask 2:

【Task Instructions】

Based on the Chinese word provided by the user, linking10 standard Chinese Mandarin Homophones or Paronyms. Requirements are as follows:

1. Arrange in descending order of pronunciation similarity, and only output 10 words;

2. All words must conform to modern Chinese word-formation rules and be common vocabulary;

3. Ensure no repetition among the 10 words;

4. Provide no explanations or additional information;

5. Attach pinyin notation to each word, indicating its pronunciation.

【Example】

Input word: '楚升(chu3 sheng1)' Output: 出生(chū shēng), 畜生(chù shēng), 出声(chū shēng), 初审(chū shěn), 出身(chū shēn), 出神(chū shén), 处身(chǔ shēn), 获胜(huò shèng), 初中(chū zhōng), 斗争(dòu zhēng)

【Input】

Input word: '{item}'

Output:

Subtask 3:

【Task Instructions】

Based on the underlined sentence provided by the user, infer the 3 most likely Chinese words that could appear in the underlined part. Requirements:

1. Only output three possible reasonable words, with no explanations or extra information.

2. The words must: fully conform to the contextual semantics, maintain grammatical correctness, and naturally collocate with the preceding and following words.

【Example】

Input sentence: 周末，我们全家一起去博物馆_____，了解了许多历史知识。

Candidate words: 参观

【Input】

Input sentence: {item}

Candidate words :

Figure 10: This figure showcases the designed prompt in the three subtasks.

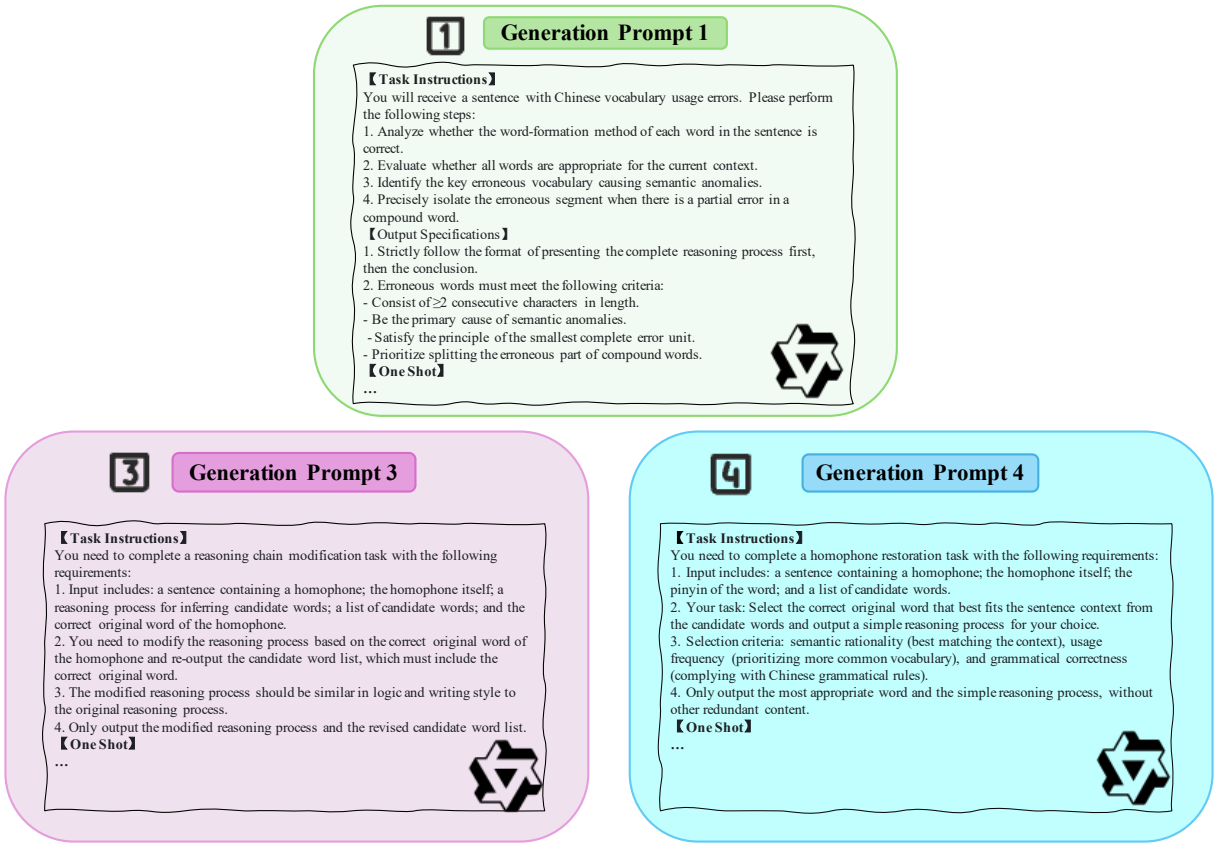

Figure 11: This figure presents the prompt structure used for Qwen2.5-72B-Instruct as a three-step reasoning generator, where each prompt follows a CoT + One-Shot format. Due to space constraints, the full content is omitted, but the detailed One-Shot examples can be found in Figure 12, where each prompt corresponds to each step. For clarity, the prompt is displayed in English, though the actual implementation used Chinese templates.

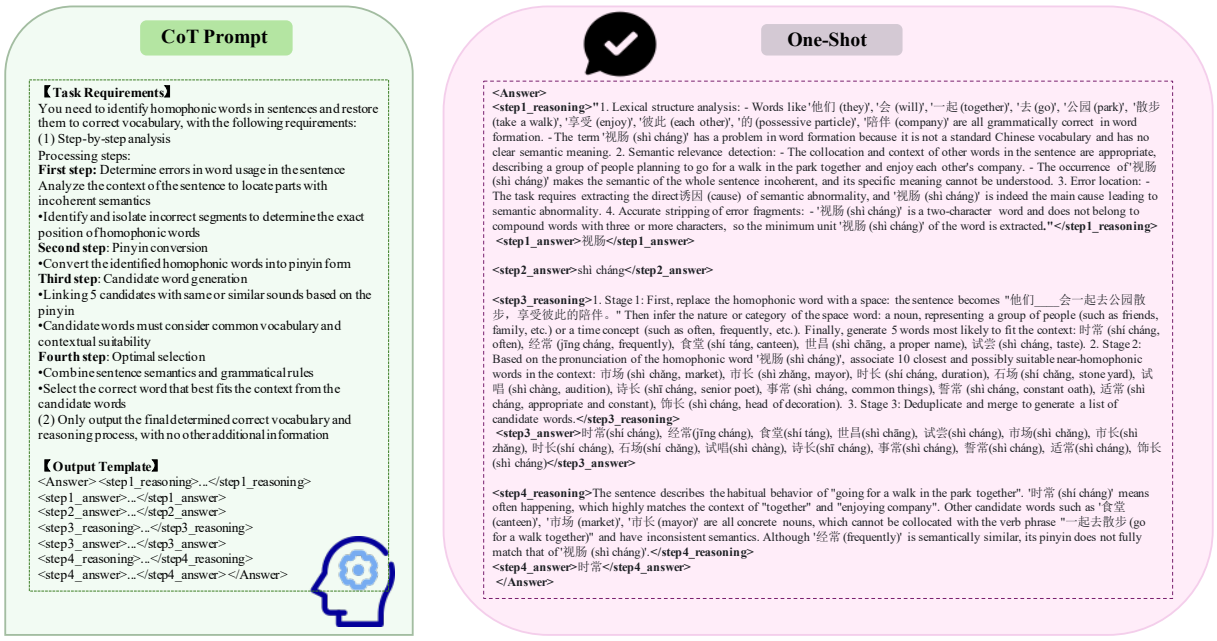

Figure 12: This figure shows the prompts of CoT used for our training and testing. Additionally, we provide a reasoning example (within the pink box) to assist the base model in understanding the task. The original prompt and example are Chinese, the given English version is for readers' understanding-friendly.

(2-4 characters), and (3) select words with POS tags a (adjective), d (adverb), n (noun), p (preposition), or v (verb). This pipeline yielded a curated dataset of 5,596 lexical entries. Finally, we employed Qwen2.5-72B-Instruct (Qwen et al., 2025) to generate contextual sentences containing each target word through prompt-based synthesis. The sentences were manually verified and corrected by three native speakers.

G.2 Synthetic Dataset

The synthetic dataset was specifically constructed for SFT and RL training purposes. The existing related Chinese datasets predominantly focus on homophones with identical syllable structures, while lacking coverage of paronyms (similar but non-identical syllables). So we systematically generated synthetic data by modifying the 5,596 sentences from subtask 3, replacing original words with homophones or paronyms across varying degrees of phonetic similarity.

This process involved four steps: (1) Calculated *fuzzywuzzy* similarity between each word and all entries in the *Modern Chinese Common Lexicon*; (2) Followed our 5:2:1:1:1 ratio, we divided all words in the dataset into five segments. Words from each segment were then matched with ten candidate words from five distinct similarity tiers[13]. Then we employed the same IPA-based method as in the subtask 2 dataset to re-rank the candidates and selected the word with the closest phonetic resemblance to the original - this word will be the final target word; (3) We replaced the optimal words with non-existent but phonetically identical pseudo-words in Chinese; (4) In the contextual sentence, we replaced original words with their revised target words, generating the final triplet structure (target word, original word, contextual sentence) as specified in Section 3.1's dataset schema.

[13] 50%:[100], 20%:[90-100), 10% each:[80-90)/[70-80)/[60-70).

Pseudo-code for Fuzzy *pinyin* Similarity

Input: $\begin{cases} \text{Original Text} & T \\ \text{Target Database} & H \end{cases}$

Output: Ranked words and similarity scores

Procedure:

1. Phonetic Conversion:
   $(py_T, t_T) \leftarrow \text{TOPY}(T)$
   $(py_H, t_H) \leftarrow \text{TOPY}(H)$
2. Fuzzy Matching:
   $\text{SimS} \leftarrow \text{FUZZY\_RATIO}(py_T, py_H)$
3. Variant Classification:
   $V \begin{cases} Homo & \text{if SimS} = 100 \wedge t_T = t_H \\ PAR & \text{if } 60 \leq \text{SimS} \\ Others & \text{otherwise} \end{cases}$
4. Variant Filtering:
   $V_1 \leftarrow \{v \in H \mid \text{Homo, exact pinyin match}\}$
   $V_2 \leftarrow \{v \in H \mid \text{PAR, tonal variations}\}$
   $V_3 \leftarrow \{v \in H \mid \text{Others, low similarity}\}$
5. Return:
   Rank V by SimS in descending order

Table 4: *Fuzzywuzzy* similarity calculation based on *pinyin* using thresholds to figure out target words in given data.

Pseudo-code for Panphon-based Phonetic Distance

Input: $\begin{cases} \text{Pinyin}_1 & p\_t_1 \\ \text{Pinyin}_2 & p\_t_2 \end{cases}$

Output: Normalized Similarity $S \in [0, 1]$

Procedure:

1. **Phoneme Alignment**:
   Align p_t_1 and p_t_2 using IPA segmentation
2. **Panphon Distance**:
   $D \leftarrow \text{panphon.distance}(p\_t_1, p\_t_2)$
   (Weighted feature edit distance)
3. **Similarity Conversion**:
   $S \leftarrow 1 - \frac{D - \min(D)}{\max(D) - \min(D)}$ *(Normalized to [0,1])*

Table 5: Phonetic similarity computation using Panphon's distance method. *Pinyin* was directly input with the spelling like ō and transferred into IPA to capture the articulation of sounds.