Reasoning or Memorization? Investigating LLMs' Capability in Restoring Chinese Internet Homophones

Jianfei Ma* and **Zhaoxin Feng*** and **Emmanuele Chersoni**
and **Huacheng Song** and **Zheng Chen**
Chinese and Bilingual Studies, The Hong Kong Polytechnic University
Computer Science and Engineering, Hong Kong University of Science and Technology
{jian-fei.ma,zhaoxinbetty.feng,huacheng.song}@connect.polyu.hk,
emmanuele.chersoni@polyu.edu.hk,
zchenin@connect.ust.hk

Abstract

Chinese homophones, prevalent in Internet culture, introduce rich linguistic twists to challenging language models. While native speakers disambiguate them through phonological reasoning and contextual understanding, the extent to which LLMs can effectively handle this task remains unclear, as does whether they rely on similar reasoning processes or merely memorize homophone-original word pairs in training.

In this paper, we propose **HomoP-CN**, the first Chinese Internet homophones dataset including systematic perturbations testing for evaluating LLMs' homophone restoration capabilities. With the benchmark, we investigated the influence of semantic, phonological, and graphemic features on LLMs' restoration accuracy, measured the memorization reliance levels of each model during restoration through consistency ratios under controlled perturbations, and assessed the effectiveness of various prompting strategies, including contextual cues, *pinyin* augmentation, few-shot learning, and thought-chain [1].

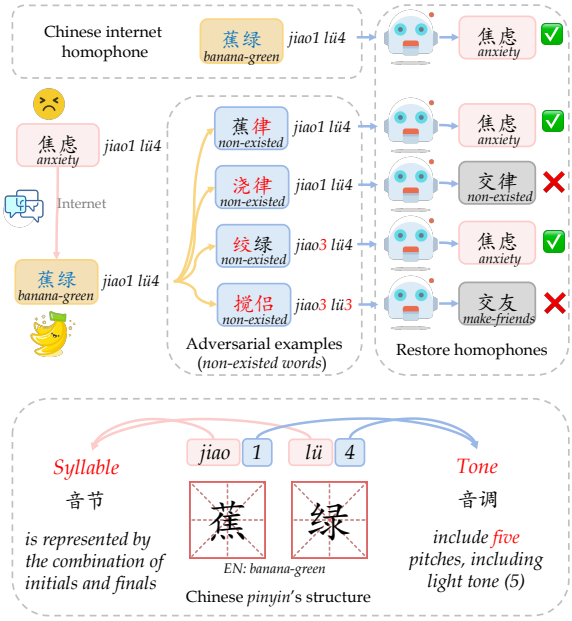

Figure 1: The upper figure illustrates an example of the homophonic word and its different adversarial perturbations in **HomoP-CN** dataset. The bottom figure demonstrates the structure of Chinese *pinyin*, which encompasses a syllable and a tone.

1 Introduction

Homophonic wordplay in Chinese Internet culture creatively utilizes phonological similarities between characters to construct new words and layered semantic meanings (Zhang et al., 2019). For example, the homophone "蕉绿" (*jiao1 lü4*, "banana-green") replaces the original word "焦虑" (*jiao1 lü4*, "anxiety"), reconfiguring a negative emotion into a playful and lighthearted expression. Unlike English puns, which rely on intralingual homophony (e.g., "a good pun is its own reword/reward") (Xu et al., 2024), Chinese homophonic wordplay creatively substitutes characters with similar pronunciations within the logographic writing system.

Recent advances in natural language processing (NLP), particularly through large language models (LLMs), have demonstrated substantial progress in disambiguating English homophones (Proietti et al., 2024; Xu et al., 2024; Mizrahi et al., 2024). However, due to the high homophone density in Chinese *pinyin*(e.g., *shi4* mapping to dozens of characters such as 是/事/市) and tonal complexity (identical syllables with different tones convey distinct meanings, e.g., *ma1*妈/*ma2*马/*ma3*码/*ma4*骂), LLMs encounter greater challenges in comprehending Chinese homophones than English.

Previous research has explored Chinese homophones in NLP tasks such as spelling correction (Liu et al., 2025, 2024; Li et al., 2024; Baluja, 2025), offensive language detection (Xiao et al., 2024), and humor generation (Xu, 2024). Nev-

* represents these authors contributed equally to this work.
[1]Our code and data are released at: https://github.com/sdmjf/Chinese_homophone_restoration_LLM.

Proceedings of the 3rd Workshop on Towards Knowledgeable Foundation Models (KnowFM), pages 120–139
August 1, 2025

ertheless, there is no systematic study on LLMs' ability to understand and restore Chinese homophones, which is crucial for practical applications such as improving LLMs' ability to understand social media text and identifying offensive content. For instance, Chinese netizens may replace "太贱 (too mean)" with the same pronunciation "肽键 (peptide bond)" to use a non-offensive biological term conveying discriminatory and offensive content (Xiao et al., 2024).

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, we propose the following research question: *How do LLMs perform in homophone restoration? Is this capability of LLMs driven by human-like reasoning through phonological similarity, or simply stem from memorization of homophone-original word pairs in pretraining? Additionally, can strategies like contextual information or providing pinyin[2] to enrich prompts enhance LLM performance in restoration?*

In this work, we comprehensively explored LLMs' effectiveness and enhancement in Chinese Internet homophone restoration by utilizing our **HomoP-CN** dataset. First, we analyzed the restoration capacity of LLMs by considering the differences between the original words and the homophones from semantic, phonological, and graphemic perspectives. Second, drawing inspiration from Xie et al. (2024), we designed a set of adversarial variations as perturbations to quantify the extent of memorization, as shown in Figure 1. Finally, we delved into the role of different prompting strategies, including context cues, *pinyin*-augmentation, few-shot, Chain-of-Thought (CoT) (Kojima et al., 2022), Memory-of-Thought (MoT) (Li and Qiu, 2023) in this task.

Our results demonstrate that LLMs exhibit substantial variation in restoring Chinese Internet homophones, with model scale emerging as a critical factor: larger models achieve reasoning-based restoration while smaller ones depend predominantly on memorization. This performance gap is further modulated by semantic, phonological, and graphemic disparities between original words and their homophone counterparts, which systematically affect both restoration accuracy and memorization dependence. While contextual cues, few-shot learning, and thought-chain strategies (CoT/MoT) prove effective for performance enhancement, *pinyin* augmentation shows limited utility. These findings provide valuable insights into LLMs' robustness in handling intralingual and user-generated content in Internet contexts.

2 Related work

2.1 Chinese Homophones

English homophones are words with distinct meanings that share the same pronunciation but differ in spelling. (HarperCollins, 2023). Similarly, in Chinese, homophones refer to a linguistic phenomenon where different words or phrases have similar or identical pronunciations (i.e., sharing the same or similar *pinyin*) but are represented by different Chinese characters[3]. On the Internet, homophones are frequently employed to substitute for or allude to the meanings of certain original words, often serving humorous or euphemistic purposes in communication (Xiao et al., 2024; Xu, 2024).

Current research on the ability of LLMs to comprehend Chinese homophones remains limited and is scattered across various NLP tasks. In spelling correction, LLMs face bottlenecks in coordinating phonological, graphemic, and semantic features when distinguishing between homophones (Liu et al., 2025, 2024; Li et al., 2024). For offensive language detection, LLMs demonstrate reduced effectiveness in identifying homophone-disguised toxic content, revealing vulnerabilities in understanding when confronted with phonological interference (Xiao et al., 2024). Additionally, LLMs exhibit challenges in semantic reasoning for humor generation involving homophones (Xu, 2024).

2.2 Language Perturbation

Researchers have proposed a wide range of perturbation techniques to explore the vulnerabilities of NLP models in adversarial scenarios, particularly

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

[3] Chinese internet homophones include both perfect homophones and near-homophones (paronyms). Many of these words do not actually exist in standard Chinese, like "蕉绿" ("banana-green"). This encompasses: 1) Characters with identical pronunciation (same syllable + tone); 2) Characters with the same syllable but different tones; 3) Similar-sounding syllables where some phonetic feature differs (e.g., z/zh distinction between apical anterior and posterior consonants, ignoring tone differences).

through replacements or insertions at the character, word, and sentence levels (Alzantot et al., 2018; Jin et al., 2020; Ribeiro et al., 2020; Zhang et al., 2020; Garg and Ramakrishnan, 2020).

Recent studies have explored Chinese adversarial attacks through various language-specific perturbations, such as synonym substitution (Su et al., 2022), phonological and glyph swaps (Liu et al., 2023; Wang et al., 2024), and emoji replacement (Xiao et al., 2024). However, no studies have yet focused on the lexical perturbations for the Chinese homophone restoration task. Our work addresses this gap by introducing the **HomoP-CN** dataset, which provides different adversarial examples tailored to the unique characteristics of Chinese homophones.

2.3 Memorization in LLMs

The memorization capabilities of LLMs have been extensively studied across multiple domains, including copyright (Karamolegkou et al., 2023; Wei et al., 2024), logical reasoning (Xie et al., 2024), and performance on knowledge-intensive tasks (Hartmann et al., 2023). Previous studies have demonstrated that LLMs are capable of memorizing portions of their training data (Tirumala et al., 2022; Carlini et al., 2022).

In this paper, we focus on quantifying the extent of memorization in LLMs when performing the homophone restoration task. Inspired by Xie et al. (2024), we designed a set of adversarial variations to quantify the extent of memorization within a controlled setting: significantly worse performance on variants versus original homophones and suggests greater reliance on memorization[4].

3 Methodology

3.1 Problem Definition

Let $D = \{(X, Y)\}$ denote a dataset where each consists of a homophone X and the corresponding original word Y. The task of LLM is to analyze X and select a word $\hat{Y}$ which is most likely to be the original word Y. Formally, the output can be represented as:

$$\hat{Y} \sim \pi_\theta(X), \tag{1}$$

The goal of LLMs is to ensure $\hat{Y} = Y$, meaning that the LLMs correctly restore the target word. In this study, we use accuracy to represent the model's performance in the task of restoring homophones.

3.2 Dataset Construction

The **HomoP-CN** dataset involves extracting homophonic words from mainstream Chinese social media platforms as the control set, followed by a multi-faceted process of categorization and conversion. This enables a systematic comparison of the performance of LLMs across various dimensions and factors. Further details are outlined below.

3.2.1 Data Collection and Categorization

Given the prevalence of homophones particularly in creative and flexible online contexts, this study sourced target homophones from two mainstream Chinese social media platforms, namely, Weibo and Tieba[5]. After reviewing a random collection of user-generated posts and comments from these platforms first, spanning the period from 2010 to 2025 (before the data cutoff in March), a total of 365 highly frequent and representative homophonic words were filtered out by three native Chinese speakers with consensus, who also provided the original word and *pinyin* for each homophone. Besides, to explore the potential impacts of contexts, we augmented the homophones into sentences with sufficient contextual information by which humans can accurately infer their original words. All context sentences were generated by the DeepSeek-V3 model (DeepSeek-AI, 2024) with the prompts shown in Appendix A.4 and then validated by three native speakers (Appendix A.2).

Upon this preliminary dataset, we further grouped all homophones in line with three distinct taxonomies for fine-grained evaluation of LLMs' performance concerning their different semantic, phonological, and graphemic properties. The semantic categorizations were completed by three native speakers based on instruction guidance (Appendix A.2), and those at phonological and graphemic aspects were sorted through automated annotated methods by comparing the distinction in the form of the *pinyin* and characters in homophones and their origins (Appendix A.3). Examples are displayed in Figure 2.

[4]Borrowing intuition from human behavior: Students preparing for exams might not fully grasp underlying principles due to constraints. Yet, they can answer memorized exact questions correctly. A key trait of such memorization is high accuracy on identical questions but poor performance on slightly modified, similarly difficult ones.

[5]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.

Original word	Homophone	Original word pinyin	Homophone pinyin	Semantics	Phonology	Graphemics	Variant 1	Variant 2	Variant 3	Variant 4
什么 *Everything*	神马 *God-horse*	*shen2 me5*	*shen2 ma3*	1	4	3	神玛 *shen2ma3*	什玛 *shen2ma3*	神吗 *shen2ma1*	审妈 *shen3ma1*
Context sentence: 神马都是浮云。*(Everything's just a puff piece.)*										
悲剧 *Tragedies*	杯具 *Cup*	*bei1 ju4*	*bei1 ju4*	2	1	3	碑具 *bei1ju4*	碑据 *bei1ju4*	贝具 *bei4ju4*	倍菊 *bei4ju2*
Context sentence: 杯具总是让人心情沉重。*(Tragedies always make people feel heavy-hearted.)*										
压力 *Pressure*	鸭梨 *Ya pear*	*ya1 li4*	*ya1 li2*	2	2	3	鸭黎 *ya1li2*	鸦黎 *ya1li2*	鸭莉 *ya1li4*	讶利 *ya4li4*
Context sentence: 鸭梨好大，我想去散步放松一下。*(I'm under so much pressure, and I want to go for a walk to relax.)*										

Figure 2: Data examples from our dataset. The numbers in the *Semantics*, *Phonology*, and *Graphemics* columns indicate the categories of homophones based on their differences from the original words in these three aspects, while *Variants* are adversarial perturbations. For detailed descriptions, refer to Section 3.2.

- **Semantic taxonomy** The first taxonomy labeled target homophones into two groups based on their semantic features on word level: 1) those are existing words and have meaning on their own and 2) those are pseudo words that are inherently meaningless.

- **Phonological taxonomy** Based on the phonological features of homophones, they were further grouped into: 1) homophones sharing matching syllables; 2) those with matching syllables but differing tones; 3) those with matching tones but differing syllables; and 4) those with differing syllables and tones, when compared to their corresponding original words.

- **Graphemic taxonomy** Refer to the difference in typing form and length of characters, all homophones were categorized into three groups, covering: 1) homophones with fewer characters (partially same or completely different) than their corresponding original words; 2) homophones sharing the same length and partially same characters with their origins; and 3) homophones with the same length but completely different characters compared to corresponding original words[6].

3.3 Task Formulation

This section delineates the design of progressive tasks aimed at evaluating the capabilities of LLMs in homophone restoration and uncovering the underlying patterns governing their performance. Considering the results from the ablation study in Appendix B.1, we selected Chinese as the language of prompts, whose detailed examples are presented in Appendix B.2.

Restoring Capability Under Zero-shot

To investigate whether popular LLMs can identify the profound relationships between pronunciations and meanings for Chinese characters in homophones, and the extent to which they can do so, we provided basic zero-shot prompts to each LLM, instructing them to restore the original forms from specific homophones. This task was conducted under three setups introduced in Section 3.2.1 to examine whether the semantic, phonological, and graphemic properties of homophones pose different challenges to LLMs and whether LLMs exhibit varying sensitivity to these properties. The metric of accuracy was employed to quantify performance by calculating the percentage of correct answers.

Patterns Behind Homophone Restoration

What follows the assessment of global restoration performance among LLMs is whether their capabilities are predominantly grounded on memorization of training data or reasoning derived from phonological similarity. To pursue this, we employed four adversarial variants as described in Section 3.4 with basic zero-shot prompts for perturbation. Besides, we define the Consistency Ratio (*CR*) to measure how robustly a model restores homophone variants. For each correctly restored case from basic homophones, we count how many of its four variants are also correctly returned to the original form, then average this count across the number of all restored cases from basic homophones. The final *CR* score (between 0 and 1) is obtained by normalizing this average against the maximum possible correct variants per homophone. Higher *CR* indicates less reliance on memorization and more on reasoning. Formally, *CR* can be represented as:

[6]Since each Chinese character covers a single syllable, the difference in character numbers between a homophone and its origin reflects elision or assimilation in their pronunciation.

Model	*Homophone*	*Variant1*	*Variant2*	*Variant3*	*Variant4*	*Variants Avg*
Llama3.1-8B	0.052	0.025	**0.030**	0.022	0.011	0.022
Qwen2.5-7B	0.216	**0.099**	0.060	0.082	0.019	0.065
OpenAI o3-mini	0.622	**0.422**	0.337	0.386	0.345	0.373
Deepseek-R1	0.833	**0.636**	0.515	0.537	0.370	0.514

Table 1: Results of the basic prompt experiments, including the accuracy of homophones, that of four types of adversarial variants, and the average value of variants. The best results among the variants are **bolded**, and the second-best results are underlined.

$$CR = \frac{1}{|D_{\mathrm{C}}|} \sum_{X \in D_{\mathrm{C}}} \left(\frac{1}{4} \sum_{i=1}^{4} \mathbb{I}[f(X_i') = Y_i'] \right) \quad (2)$$

Where $D_{\mathrm{C}} = X \in D \mid f(X) = Y$ (set of successfully restored homophones), X_i' denotes the i-th variant of homophone X, Y_i' is the correct orginal word for variant X_i', and $\mathbb{I}[\cdot]$ is the indicator function.

Impacts of Context Cues and Other Strategies

In our final exploration, we investigate the impacts of several related knowledge and prompt strategies on LLMs' performance in homophone restoration. It is widely acknowledged that humans typically infer the meanings of homophones based on contextual cues at first glance (Xu et al., 2024). Hence, we first examine the effects of contextual information. The context sentences created in Section 3.2.1 were integrated into the basic zero-shot prompts, and the results were compared with those from the basic prompt. Improved performance indicates that contextual information positively contributes to homophone restoration, while degraded performance suggests the opposite.

Building on this exploration, we further investigate the impact of additional strategies, including: 1) Few-shot prompts: Provide examples to guide the model; 2) *Pinyin* annotations: Supply *pinyin* for sentences and homophones; 3) CoT: Encourage step-by-step reasoning; and 4) MoT: Leverage memory-enhanced reasoning, to provide comprehensive insights to this task.

3.4 Data Perturbation

To examine the underlying patterns of LLMs in restoring homophones, we created adversarial scenarios against the control homophones by introducing semantic, phonological, and graphemic perturbations through character modifications, as illustrated in Figure 1 and Figure 2.

We utilized a well-compiled dictionary[7] including 2,715 common Chinese characters with *pinyin* spellings, to enable automatic character retrieval and replacements. Using this dictionary, we introduced four types of adversarial variants with incremental distances away from the control homophones by replacing one character (or all characters) of the control homophone with a different character (or some different characters) sharing the same *pinyin* or the same syllables but differing in tones (Appendix A.5).

3.5 Model Selection

Models applied in current study include Qwen 2.5-7B (Qwen et al., 2025), Llama 3.1-8B (Grattafiori et al., 2024), OpenAI o3-mini[8] and Deepseek-R1 (DeepSeek-AI et al., 2025). Among these, the former two are open-sourced, while the latter two are not. These models rank at the top of current leaderboards and demonstrate remarkable performance across a diverse set of tasks, including reasoning, by leveraging extensive memorization capabilities developed during the pre-training phase (Zhang et al., 2024; Prabhakar et al., 2024). For all models, the temperature is set to 0, and other configurations are applied as default.

4 Results and Analyses

4.1 *Can LLMs Restore Chinese Internet Homophones to Original Words?*

As shown in Table 1, LLMs show significantly distinct performance in restoring homophones. Llama 3.1 and Qwen 2.5 show an overall weak performance. OpenAI o3-mini demonstrates superiority by outperforming the first two models, and Deepseek-R1 achieves the best performance with outstanding accuracy, showcasing its robustness correspondingly. The performance differences may

[7] https://github.com/5hwb/sort-hanzi-in-pinyin-order/
[8] https://openai.com/index/openai-o3-mini/

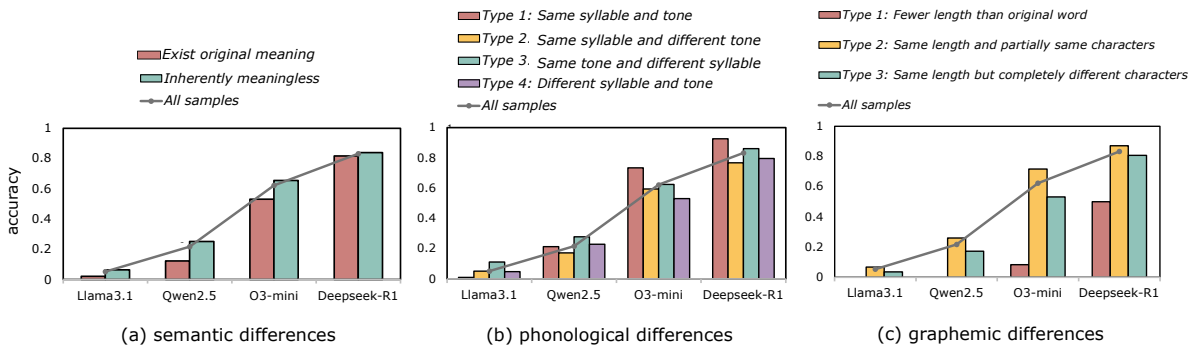

Figure 3: The impact of semantic, phonological, and graphemic disparities between homophones and their original words on LLMs' homophone restoration performance.

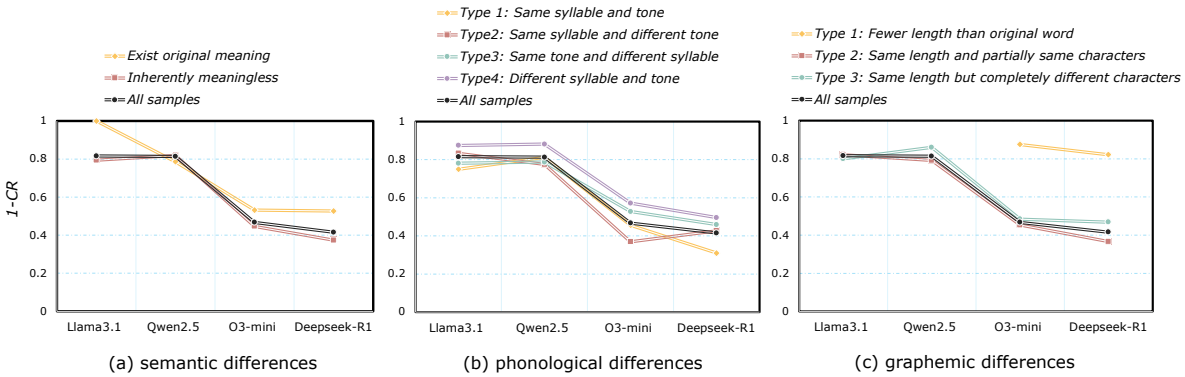

Figure 4: The influence of semantic, phonological, and graphemic differences between homophones and original words on the level of memorization dependence during homophone restoration. A higher (1-*CR*) value on the y-axis indicates greater reliance on memorization by LLMs.

stem from variations in training data and model size: Deepseek-R1 used more extensive Chinese corpora in training, outperforming the other models with the same scale. OpenAI o3-mini and Deepseek-R1, with larger parameter sizes and stronger inference capabilities, excel in this complex linguistic task than smaller models.

Also, we systematically categorized homophones based on differences between homophones and original words in terms of semantics, phonology, and graphemics to explore how these characteristics influence LLMs' ability to restore homophones. The results are shown in Figure 3.

For the semantic dimension, homophones were divided into two categories: those whose original word-level meanings exist and those that do not exist. The results in Figure 3 (a) reveal that all LLMs exhibit stronger restoration capabilities for homophones without existing meanings, suggesting that the inherent semantics in homophones may interfere with restoration, especially for small models.

For phonological differences, Figure 3 (b) shows significant accuracy differences in LLMs' homophone restoration across four types. Type 1 (consistent syllables+tones) outperformed others, except in large-parameter LLMs. Type 3 (same tone+different syllables) worked well in small models. Both types highlight the essentials of pinyin syllables and tones in homophone restoration. However, when comparing the performance of Type 2 and Type 3 to Type 1, it is emphasized that the same tone can benefit more than syllables in large models. Small models are highly dependent on the same tone, and syllables even negatively affect the accurate prediction of original words.

For the graphemics dimension, Figure 3(c) shows that Llama 3.1 and Qwen 2.5 completely fail to restore Type 1 homophones (shorter characters replacing original words, e.g., "酱紫" replaces "这样子", meaning "like this"). Even large models perform worst on Type 1 homophones, indicating that LLMs struggle most with pronunciation elision. In contrast, LLMs excel at Type 2 and 3 homophones, which have the same length with

partial or total character substitutions, highlighting their sensitivity to word length and subtle surface graphemic changes.

4.2 *Reasoning or Memorization?*

To determine whether LLMs restore Chinese homophones primarily through memory or reasoning, we conducted experiments using four types of adversarial variants. Results are presented in Figure 4:

Deepseek-R1 and OpenAI o3-mini exhibit significantly less reliance on memorization compared to the other two models, likely attributable to their much larger scale and enhanced reasoning capabilities. Notably, Llama 3.1 demonstrates near-total reliance on memorization when the homophone carries their inherent semantic meanings.

Figures 4 (a) and (b) demonstrate that LLMs exhibit increased reliance on memorization under two conditions: 1) when homophones retain original semantic meanings, or 2) when phonological divergence between homophones and target words grows larger. Graphically, Figure 4 (c) reveals significantly stronger memorization dependence when homophones contain fewer characters than their corresponding original words.

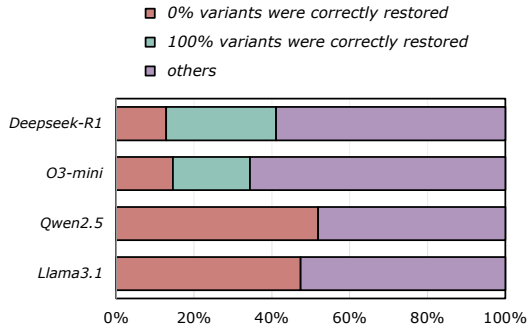

Figure 5: Percentage of different *CR* value homophones in the four LLMs.

We computed the *CR* for each successfully restored homophone, where *CR* = 1 indicates perfect variants restoration (100% accuracy) and *CR* = 0 denotes complete failure, as shown in Figure 5. Our results align with the pattern in Figure 4: smaller models demonstrate notably poorer performance on perturbation data compared to larger models.

Furthermore, based on the experimental results from Deepseek-R1 and OpenAI o3-mini, we selected homophones with different *CR* value, comparing their distributions across: 1) original word frequency[9] and 2) homophone-original word

[9]Calculated by Python library *wordfreq*, available at link.

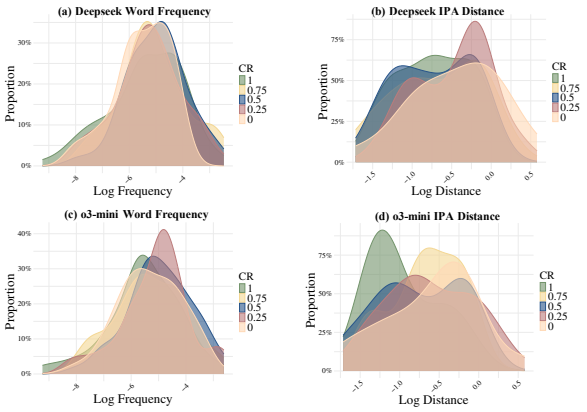

Figure 6: Comparison of homophone properties across *CR* values in Deepseek-R1 and OpenAI o3-mini. Analyzed distributions include: 1) original word frequency (log-scaled), and 2) IPA-based phonological distance between homophones and original words. Higher *CR* indicates less reliance on memorization.

phonological distance based on the *International Phonetic Alphabet* (see Appendix B.3). Intuitively, we hypothesize that LLMs rely more heavily on memorization when processing: 1) those derived from high-frequency original words (leveraging their prevalence in training data), and 2) those exhibiting substantial phonological divergence from their original words.

Results are shown in Figure 6. Contrary to our hypothesis, the original word frequency showed little correlation with memorization dependence during homophone restoration. Instead, phonological divergence between homophones and original words emerged as a more dominant factor (consistent with what we observed in Figure 4), particularly in OpenAI o3-mini.

4.3 *Can Contextual Cues Enhance Homophone Restoration in LLMs?*

The basic assumption for examining the effect of contextual information in homophone restoration is that, as for humans, contextual information can restrict and redirect potential choices in a more narrow range, facilitating accurate predictions. Thus, context-enhanced prompts (Appendix B.2) were employed to assess the role of context in improving LLMs' restoration performance.

As shown in Table 2 and Figure 7, the context-enhanced prompt can improve LLMs' restoration ability on Chinese homophones. Context imposes constraints, guiding LLMs to generate restored words relevant to the given semantic and pragmatic environment. As presented in Table 2, for all four

Model	*Homophone*	*Context*	*+Fewshot*	*+Pinyin*	*+CoT*	*+MoT*
Llama3.1-8B	0.052	0.058	0.156	0.036	0.099	**0.164**
Qwen2.5-7B	0.216	0.293	0.356	0.269	0.277	**0.400**
OpenAI o3-mini	0.622	0.723	0.732	0.723	0.718	**0.762**
Deepseek-R1	0.833	0.896	**0.910**	0.896	0.871	**0.910**

Table 2: Results of the context-enhanced prompt experiments. The best results among the *Fewshot*, *Pinyin*, *CoT*, and *MoT* are **bolded**, and the second-best results are underlined.

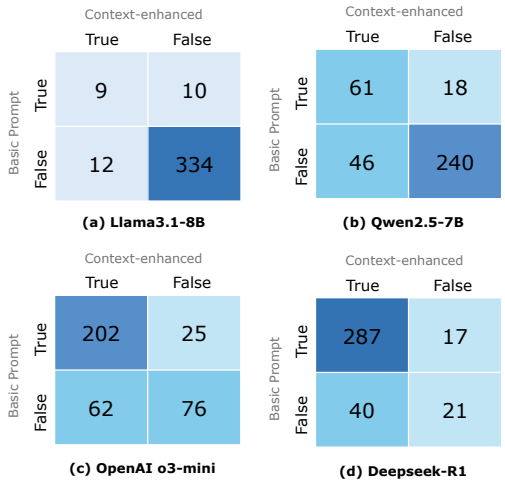

Figure 7: Comparison of the basic prompt and context-enhanced prompt experiments' results.

LLMs, contextual information can evidently improve their performance of restoring homophones into their original words (see the increased accuracy from *Homophone* column to *Context* column). However, in-depth results in Figure 7 uncover that this improvement is not universal. In other words, some cases correctly restored in the basic prompt experiment would be incorrectly handled after adding context. Specifically, for example, in Llama 3.1, 10 such cases can be observed (see upper right block), a phenomenon also seen in other models. This suggests that contextual information does not consistently impose a positive effect on each Chinese homophone for restoration and can sometimes disrupt comprehension or impair memorization in LLMs.

4.4 *Can Other Strategies Impact Homophone Restoration in LLMs?*

This study further examines if other strategies can enhance restoration ability. Table 2 summarizes the contributions of the different strategies.

Few-shot learning and MoT prompts can significantly enhance the restoration performance by presenting human-annotated examples to LLMs. Examples from few-shot learning can reveal linguistic patterns of homophones to LLMs, while MoT prompts explicitly provide human reasoning logic and *pinyin*-based knowledge. This enables LLMs to adopt these reasoning strategies, further improving their restoration capabilities.

***Pinyin* augmented prompts** result shows that LLMs have difficulty in explicitly adapting this knowledge alone to assist homophone restoration. This suggests that their orthography training may limit their effective leverage of *pinyin*.

CoT prompts realized various performance fluctuations among models. Specifically, Llama 3.1 improves with CoT, while Qwen2.5, OpenAI o3-mini, and Deepseek-R1 show declines. This discrepancy may arise from their default reasoning strategies. This task requires simultaneous pinyin and contextual information rationale. Without effective guidance for basic CoT prompts, Qwen2.5, OpenAI o3-mini, and Deepseek-R1 are prone to follow the default think flow, leading to errors in reasoning. In contrast, Llama 3.1 benefits from CoT as it compensates for its default lack of reasoning emphasis, improving restoration accuracy.

5 Conclusion

In this study, we present the first Chinese Internet homophones dataset with language perturbations to evaluate LLMs' restoration capabilities and their reliance on memorization. Our results show that LLMs exhibit significant differences in restoring homophones: larger models rely more on reasoning, while smaller ones depend on memorization. Performance variations are further influenced by semantic, phonological, and graphemic differences between original words and homophones, systematically affecting accuracy and memorization dependence. Although strategies like contextual cues, few-shot learning, and MoT improve performance, *pinyin*-based augmentation unexpect-

edly failed to enhance restoration. These findings shed light on LLMs' robustness with intralingual and user-generated online content.

Ethics Statement

We do not foresee any ethical risks related to our research.

Limitations

This study quantifies the extent of memorization in LLMs' restoration of Chinese homophones, though the underlying mechanisms of restoration remain unclear. A limitation is the use of DeepSeek-Chat to generate context sentences, which, despite human proofreading and optimization, may still impact experiments involving contextual prompts[10].

Additionally, our study is confined to four models (Llama3.1-8B, Qwen2.5-7B, OpenAI o3-mini, and Deepseek-R1), and results may vary with other models. Future work should expand to diverse languages and models to validate and refine these findings.

Moreover, character co-occurrence and character frequency are likely to influence the memorization and reasoning processes of LLMs during homophone restoration. Currently, there are no up-to-date datasets that incorporate Chinese Internet homophones along with data on character co-occurrence and character frequency. Future research efforts are expected to concentrate on collecting such data, with the aim of further exploring the impact of character co-occurrence and character frequency on homophone restoration.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Ashwin Baluja. 2025. Text is not all you need: Multimodal prompting helps LLMs understand humor. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 9–17, Online. Association for Computational Linguistics.

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.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

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.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu,

[10]To address this concern, we include supplementary real-world data evaluations in the Appendix C.1, where all LLMs demonstrate consistent performance trends between authentic and synthetic data, validating the reliability of our main synthetic-data findings.

Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Siddhant Garg and Goutham Ramakrishnan. 2020. BAE: BERT-based adversarial examples for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6174–6181, Online. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzm á n, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur elebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, V í tor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis,

Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

HarperCollins. 2023. *Collins English Dictionary: Complete and Unabridged*, 14th edition. HarperCollins. ISBN 9780008511340, 1899 pages.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models. *Preprint*, arXiv:2310.18362.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Xiaonan Li and Xipeng Qiu. 2023. MoT: Memory-of-thought enables ChatGPT to self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6354–6374, Singapore. Association for Computational Linguistics.

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

Hanyu Liu, Chengyuan Cai, and Yanjun Qi. 2023. Expanding scope: Adapting English adversarial attacks to Chinese. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 276–286, Toronto, Canada. 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.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Akshara Prabhakar, Thomas L. Griffiths, and R. Thomas McCoy. 2024. Deciphering the factors influencing the efficacy of chain-of-thought: Probability, memorization, and noisy reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3710–3724, Miami, Florida, USA. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Simone Tedeschi, Giulia Vulpis, Leonardo Lavalle, Andrea Sanchietti, Andrea Ferrari, and Roberto Navigli. 2024. Analyzing homonymy disambiguation capabilities of pretrained language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 924–938, Torino, Italia. ELRA and ICCL.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

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

Hui Su, Weiwei Shi, Xiaoyu Shen, Zhou Xiao, Tuo Ji, Jiarui Fang, and Jie Zhou. 2022. RoCBert: Robust Chinese bert with multimodal contrastive pretraining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–931, Dublin, Ireland. Association for Computational Linguistics.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Ao Wang, Xinghao Yang, Chen Li, Bao-di Liu, and Weifeng Liu. 2024. Adaptive immune-based sound-shape code substitution for adversarial Chinese text attacks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4553–4565, Miami, Florida, USA. Association for Computational Linguistics.

Boyi Wei, Weijia Shi, Yangsibo Huang, Noah A. Smith, Chiyuan Zhang, Luke Zettlemoyer, Kai Li, and Peter Henderson. 2024. Evaluating copyright takedown methods for language models. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*. Datasets and Benchmarks Track.

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.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2024. On memorization of large language models in logical reasoning. In *NeurIPS 2024 Workshop MATH-AI: The 4th Workshop on Mathematical Reasoning and AI*.

Rongwu Xu. 2024. Exploring chinese humor generation: A study on two-part allegorical sayings. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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.

Dongyu Zhang, Heting Zhang, Xikai Liu, Hongfei Lin, and Feng Xia. 2019. Telling the whole story: A manually annotated Chinese dataset for the analysis of humor in jokes. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6402–6407, Hong Kong, China. Association for Computational Linguistics.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

Yizhuo Zhang, Heng Wang, Shangbin Feng, Zhaoxuan Tan, Xiaochuang Han, Tianxing He, and Yulia Tsvetkov. 2024. Can LLM graph reasoning generalize beyond pattern memorization? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2289–2305, Miami, Florida, USA. Association for Computational Linguistics.

A About Dataset

A.1 Human Annotation

In our paper, two aspects require annotation: whether homophones have their meanings as existing words in Chinese and whether the sentence carriers generated by Deepseek-V3 are appropriate for the target homophone.

Regarding the first task, the Chinese homophones in the dataset are assigned to three native Chinese researchers in linguistics for annotation. When their opinions are not in agreement, we adopt the annotation results of the majority. For the latter task, we invited the same annotators to make them to determine the suitability of the homophone for the given sentence. Subsequently, sentences with different opinions will be further revised until a full agreement is reached for further implementation.

A.2 Instruction for Annotators

Judgment of Homophone Inherent Meaning

- Please check each of the homophones below and determine whether they have an inherent meaning in Chinese as an existing word. For example, "压力 (pressure)" and its homophone "鸭梨 (Chinese white pear)", where "鸭梨" has its original meaning as a fruit, and this kind of homophone should be marked as "1". Another example "焦虑 (anxiety)" and its homophone "蕉绿 (banana-green, which has no independent semantic meaning in standard language)", which should be marked as "0".

- In general, if the homophone is a word with a clear semantic meaning in regular language use, mark it as "1". If the homophone is just created as a homophone and has no actual semantic meaning, mark it as "0".

The Inter-Annotator-Agreement (IAA) reaches 92.88% (Fleiss' Kappa = 0.7215). Interannotators' inconsistent cases will finalize the label by the majority choice.

Judgment of Carrier Sentences Suitability

- Please review these given sentences that carry homophones. Your task is to determine whether each sentence conforms to the functions of Chinese homophones. If the sentence is appropriate, mark it as "1". If not (such as incorrect grammar or inappropriate context), mark it as "0".

A.3 Pseudo-code for Categorization

The pseudo-code of grouping words based on phonological and graphemic features is shown in Table 3. *Pinyin* can be accessible by involving the *pypinyin* package directly to transfer the character into Chinese *pinyin*. The pseudo-code demonstrates the logic of transferring words into Chinese *pinyin* and conducting the phonological taxonomy based on *pinyin* syllable and tone distinction. The package can transfer the neutral tone into label "5".

Pseudo-code for Categorization

Input: and $\begin{cases} \text{Original Word} & ow \\ \text{Homophone} & h \end{cases}$

Output: Result r

Procedure:

$py_o r = f_{\text{char2pinyin}}(ow)$

$py_h o = f_{\text{char2pinyin}}(h)$

$(s_1, t_1) = py_o r$

$(s_2, t_2) = py_h o$

$$r = \begin{cases} 0, & \text{if } s_1 = s_2 \text{ and } t_1 = t_2 \\ 1, & \text{if } s_1 = s_2 \text{ and } t_1 \neq t_2 \\ 2, & \text{if } s_1 \neq s_2 \text{ and } t_1 = t_2 \\ 3, & \text{if } s_1 \neq s_2 \text{ and } t_1 \neq t_2 \end{cases}$$

return r

Table 3: This table demonstrates the logic of phonological taxonomy based on *pinyin* syllables and tone.

A.4 Prompt for Carrier Sentence Generation

The carrier sentences are generated from the Deepseek-V3 based on the prompt shown in Table 4. Since Deepseek-V3 may not be able to understand the meanings of certain homophonies, we first input the original words into Deepseek-V3. This step allows Deepseek-V3 to generate carrier sentences based on the original words. Subsequently, we replace the original words with their corresponding homophones. Finally, we present these sentences with replaced homophonic words to annotators for verification and modification.

A.5 Pseudo-code for Adversarial Variants Generation

The pseudo-code of homophone variants generation is shown in Table 5.

Chinese original version
/*指令*/ 你是一位中文语言专家，擅长创作简单的句子。请你根据输入中的词创作符合逻辑的句子，要求结构简单，使用场景日常。 /*示例*/ 怎么了 - > 怎么了？身体不舒服吗？ /*输入*/
English translated version
/***Instructions***/ You are a Chinese language expert and you are good at writing simple sentences. Please create logical sentences based on the words in the input. The structure should be simple and the use situation should be daily. /***Example***/ What's wrong - > What's wrong? Don't you feel well? /***Input***/

Table 4: This table demonstrates the prompt design of Context Sentence Generation task. The input language is Chinese.

Pseudo-code for Varient Generation
Input: $\begin{cases} \text{Common Chinese Characters} & C \\ \text{Custom Homophones} & H \end{cases}$
Output: Result Variants V
Procedure:
$D = C \setminus H$
$(s_1, t_1) = f_{\text{char2pinyin}}(C)$
$(s_2, t_2) = f_{\text{char2pinyin}}(H)$
$r = \begin{cases} S_1, & \text{if } s_1 = s_2 \wedge t_1 \neq t_2 \\ S_2, & \text{if } s_1 = s_2 \wedge t_1 = t_2 \end{cases}$
Variants:
$V_1 = \{v \mid p(v) = p(c) \wedge 0.5c \neq 0.5v\}$
$V_2 = \{v \mid p(v) = p(c) \wedge c \neq v\}$
$V_3 = \{v \mid 0.5p(v) = 0.5p(c) \wedge 0.5c \neq 0.5v\}$
$V_4 = \{v \mid p(v) \neq p(c) \wedge c \neq v\}$
return Variants V

Table 5: This table demonstrates the logic of categorization based on pinyin syllables, tone, and variant generation.

B About Experiments

B.1 Ablation Study

In order to explore performance fluctuations with prompts in Chinese or English, we conducted an ablation study before the formal experiment. We applied two open-source language models: Qwen2.5-7B and Llama3.1-8B. The result is shown in Figure 8. The results indicate that Qwen2.5-7B can achieve optimal performance with English prompts in limited strategy, while Chinese prompts yield better average performance. In contrast, Llama3.1-8B obtains more optimized performance when using Chinese as the prompt language. Considering the better performance of Chinese and the nature of Chinese linguistic exploration, in our main experiment, we used Chinese as the prompt language.

In details, we explored the output of Llama and its rationale when using English prompts. This approach tends to generate coding-type words, such as "\u5c0e\u9ed1\u62a8". Consequently, the performance in most tasks reaches 0 accuracy. Additionally, when using English prompts for the CoT task, Llama3.1-8B generates rationale similar to Python programming language. This information is provided to guide you in using Python code to achieve restoration. In that case, Llama also exhibits extremely weak performance.

B.2 Details of Prompts

The step-by-step investigation on LLMs' restoration of Chinese homophones requires highly structured prompts to make LLMs understand their tasks as well as avoid the performance influence by the different context information.

Basic Prompt Design

In the basic prompt, we do not give any related information but the homophone itself to instruct LLMs for restoring based on the given homophone alone. Prompts are shown in Table 6.

Context-enhanced Prompt with *Pinyin* and Few-shot learning Design

Due to context can assist in constructing meaning via the specific contextual cue offering, our study designs the context-enhanced prompt to explore its function in restoration. Additionally, the few-shot enhanced and *pinyin* are used to further examine their influence on restoration as the homophone source-target pattern and Chinese phonological spelling role are vital for this restoration task. The prompt is shown in Table 7.

CoT and MoT prompts Design

CoT and MoT can explicitly activate the rationale of LLMs by directly showing examples in prompts.

	Qwen_EN	Llama_EN	Qwen_CN	Llama_CN
Basic Prompt	0.000	0.000	0.000	0.000
Context Enhanced	0.059	0.000	0.098	0.000
Pinyin	0.059	0.000	0.039	0.020
Few-shot Learning	0.176	0.020	0.200	0.078
CoT	0.137	0.000	0.176	0.059
MoT	0.216	0.020	0.176	0.157
AVG	0.108	0.008	**0.114**	**0.052**

Figure 8: This table demonstrates the performance of different models using different languages in given tasks. Qwen_EN, Llama_EN, Qwen_CN and Llama_CN denote Qwen2.5-7B and Llama3.1-8B using English and Chinese Prompts. AVG refers to the average accuracy of specific model with one language.

Chinese original version
/*指令*/ 你是一位专业的中文语言分析专家。你会接收到一个中文谐音词作为输入内容，请准确将其还原为原本的词汇，然后只输出一个符合“原词”：“XXX”格式的JSON数据，这里的“XXX”就是你所输入的谐音词所对应的原本词汇。 **//* 输入 */** 当前的输入是：
English translated version
/*Instructions*/ You are a professional Chinese language analysis expert. When receiving a Chinese homophone word or phrase as input, accurately revert it to its original word or phrase, then only output a JSON object conforming to the format “originalWord”: “XXX”, where “XXX” represents the original word or phrase corresponding to the input homophone term. **/*Input*/** Current input is:

Table 6: This table demonstrates the basic prompt design of restoration task. The input language is Chinese.

The CoT can allow models to reason with the default chain, while the MoT can offer the human thinking chain to let models fit to restrict the chain more task-specific and similar to humans. The prompt design is demonstrated in Table 8.

B.3 Phonological Similarity Algorithm

We applied Panphon to calculate the phonological similarity between the homophone and its corresponding original word. The detailed procedure is demonstrated in Table 10. 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.

Chinese original version
/*指令*/ 你是一位专业的中文语言分析专家。你会收到含有一个谐音词的中文句子和该句子中的拼音，和该句子中的中文谐音词和谐音词的拼音作为输入内容，请准确将其中的谐音词还原为原本的词汇，直接输出且只输出一个符合“原词”：“XXX”格式的JSON数据，这里的“XXX”就是输入谐音词所对应的原本的中文词汇。输出中只能含有该json数据，而不能包含其他任何多余信息。 **/* 示例 */** （1）句子输入：不要对我人参公鸡，否则我让管理员过来处理了。 谐音词：人参公鸡 输出为：“原词”：“人身攻击”(... with two more examples) **//* 输入 */** 句子输入： 句子输入的拼音： 谐音词： 谐音词的拼音：
English translated version
/*Instructions*/ You are a professional Chinese language analysis expert. When receiving Chinese sentence containing a homophone and the *pinyin* of the sentence, as well as the homophone in the sentence and the *magenta* of the homophone as the input, accurately revert the homophone part to the original word or phrase, then only output a JSON object conforming to the format “original word”: “XXX”, where “XXX” represents the original word or phrase corresponding to the input homophone term. **/* Examples */** (1) Input sentence is: Don’t ginseng male chicken to me, or I’ll have the warden come and deal with it. Homophone is: ginseng male chicken Output is: “original word”: “personal abuse”(... with two more examples) **/*Input*/** Input sentence is: *Pinyin* of input sentence is: Homophone is: *Pinyin* of homophone is:

Table 7: This table demonstrates the context-enhanced prompt design for restoration with two additional improvement strategies. The text with this text color denote the core addition of context-enhanced prompt. The text represents *pinyin* enhanced prompt. The text refers to few-shot learning enhanced prompt examples. The input language is Chinese while English translated version is a literal translation for understanding.

Chinese original version

/*指令*/

你是一位专业的中文语言分析专家。你会收到含有一个谐音词的中文句子和该句子中的谐音词作为输入内容，请首先给出思考推理的过程，然后准确地将句子的谐音词还原为原本的词汇，最后只输出样式为“推理过程”：“XXX”,“原词”：“XXX”的JSON数据，这里的第一个“XXX”是你推理的过程，第二个“XXX”就是你所输入的谐音词所对应的原本词汇。输出中只能含有该json样式的数据，而不能包含其他任何多余信息。

/* 示例 */

句子输入：不要对我人参公鸡，否则我让管理员过来处理了。

谐音词：人参公鸡

输出为：“推理过程”：“‘人参公鸡’的拼音是[[ren2],[shen1],[gong1],[ji1]]，原词应该为‘人身攻击’，拼音是[[ren2],[shen1],[gong1],[ji1]]。这是属于完全的同音字置换形成的谐音词现象，拼音拼写（发声位置）以及音调没有发生任何变化。这个谐音词中的‘参’，‘公’和‘鸡’字属于遭到置换的字。他们分别经历将’身’替换为’参’，将’攻’替换为’公’，将‘击’替换为‘鸡’。这一现象仅改变了汉字写法，保持发音一致形成了谐音效果”,“原词”：“人身攻击”...(with two more examples)

//* 输入 */

句子输入：

谐音词：

English translated version

/*Instructions*/

You are a professional Chinese language analysis expert. When receiving a Chinese sentence with a homophone words/phrase as input: please first give the rationale, then accurately revert the word or phrase in the sentence back to original form with only output a JSON object conforming to the format “reasoning process”: “XXX”, “original word”: “XXX”, where the first “XXX” is the reasoning process you carried out, and the second “XXX” represents the original word or phrase corresponding to the input homophone term.

/* Examples */

Input sentence is: Don’t ginseng male chicken to me, or I’ll have the warden come and deal with it.

Homophone: ginseng male chicken

Output: “reasoning process”: “*Pinyin* of ‘ginseng male chicken’ is [[ren2],[shen1],[gong1],[ji1]]. Original means ‘personal abuse’. This is a homophonic phenomenon formed by complete homophone replacement, but no change in spelling or tone.‘Seng’, ‘male’, and ‘chicken’ in homophone belong to replaced words. They experienced replacing ‘body’ with ‘seng’, ‘attack’ with ‘male’, and ‘strike’ with ‘chicken’ respectively. This phenomenon only changed Chinese characters, keeping the pronunciation consistent to form homophonic effects.”, “original word”: “personal abuse”...(with two more examples)

/*Input*/

Input sentence is:

Homophone is:

Table 8: This table demonstrates the CoT and MoT prompt. The text is explict activation of LLMs’ relationale. The content represents MoT with human rationale and true case. The input language is Chinese, while English translated version is given for understanding.

Model	Real Acc	Synthetic Acc
Llama3.1-8B	0.12	0.30
Qwen2.5-7B	0.64	0.80
OpenAI o3-mini	0.93	0.90
Deepseek-R1	0.99	1.00

Table 9: This table compares LLM performance using the MoT prompt on synthetic vs. real-world sentences. **Real Acc** represents the restoration accuracy for authentic homophone-included sentences, while **Synthetic Acc** denotes the accuracy for synthetic sentences with corresponding homophones.

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 10: 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.

C Additional Results Analyses

C.1 Comparison between Sentences in Real-case and Synthetic Data

This study is constrained by its reliance on synthetic data generated by LLMs, leaving real-world cases untested. Owing to the scarcity of structured data on Chinese internet homophones, we randomly selected ten homophones and sourced corresponding sentences via an online Weibo corpus with a corpus retrieval function at link, establishing a 1:10 homophone-to-sentence mapping. These real-world sentences were then applied using the MoT strategy to validate its efficacy on synthetic datasets. The results and key distinctions are summarized in Table 9.

Although we did not test all homophone cases to calculate overall accuracy, the trends observed in real-world and synthetic sentences are consistent. This suggests that synthetic data can mirror outcomes similar to real-world data and validates the feasibility of using synthetic data in the main experiment. However, the significant discrepancy

between the two also highlights that synthetic data may not fully capture the complexity of real-world scenarios, especially affecting the small models' performance a lot.

C.2 Confusion Matrices

This section reveals all confusion matrix of comparison between context-enhanced prompts with context and few-shot learning-enhanced, context and *pinyin* syllable-enhanced, context and CoT activating, and context and MoT activating prompts. Figure 9 shows the various strategies of prompts' effects on the case level. The confusion matrix highlights that the strategies of different prompts cannot consistently enhance or decline in each case. (A case can be correctly restored in one strategy, but it may be correctly or wrongly restored in subsequent strategies.)

C.3 Rationale in Error Cases Study

This session lists the original rationale in Chinese in Table 11 and Table 12.

D Experiment Details

During the experiments, we utilize one A100 GPU with 40GB of memory. Each experiment is configured to not exceed three hours in duration.

For the reasoning tasks of Deepseek-R1 and OpenAI o3-mini, we obtain access through the official API channels provided by the respective companies. As for Qwen2.5-7B and Llama3.1-8B, we download them from the official Hugging Face website and make use of the transformer package available there to integrate them into our experimental setup.

E Error Cases Study

Results from basic prompts and enhanced strategies reveal that LLMs can only restore a subset of Chinese homophones in our dataset, underscoring the challenges they face in restoration tasks. This section empirically investigates reasons behind their limitation by analyzing the rationale contents and restored words via CoT and MoT experiments. Through additional discussion of erroneous cases, we gain deeper insights into the underlying causes of these challenges. Detailed rationales for the examples are provided in Table 12 in Appendix C.3.

We manually reviewed homophones incorrectly restored by LLMs and categorized the errors into three types: 1) **Same Meaning Restoration**: The restored homophone has the same basic meaning as the original but is incomplete; 2) **Similar Meaning with Lost Elements**: The restored homophone conveys a similar meaning but loses some semantic elements of the original; 3) **Completely Wrong Restoration**: The restored word is entirely incorrect, bearing no meaningful relation to the original.

The Type 1 example, "石乐志" (*shi2 le5 zhi4*, literally "stone-happy-ambition"), was correctly restored as "失了智" (*shi2 le5 zhi4*, "lost one's mind") in the basic prompt experiment, relying on memorization. However, with CoT involvement, it was incorrectly restored as "失智" (*shi1 zhi4*, "lose mind"), omitting the past tense marker "了". This misalignment during reasoning highlights a limitation of CoT, where LLMs overthink meanings and neglect functional elements like tense markers. In contrast, the MoT prompt, which activates memorization and emphasizes proper alignment, ensures correct restoration. This suggests that LLMs' default CoT reasoning struggles to balance content words and functional elements, sometimes prioritizing meaning over structural accuracy.

The example of "雾化女性" (*wu4 hua4 nü3 xing4*, "atomization-women") in Type 2, is a partial homophone substitution and memory-relying restored homophoneme in the basic prompt experiment. However, CoT prompts incorrectly restores it as "物化" (*wu4 hua4*, "objectify"), neglecting the component of "woman", while MoT prompts can still capture all components correctly. This proves LLMs might lose their attention by CoT in dealing with multi-word tasks and tend to put the dominant focus on some key parts during restoration.

Type 3 is illustrated by the example of "非珠牛" (*fei1 zhu1 niu2*, "Non-jewelry cow"), which could be restored into "非主流" (*fei1 zhu3 liu2*, "non-mainstream") by using either memory or reasoning, as demonstrated in the basic prompt experiment. However, when guided by CoT prompts, it is incorrectly restored as "非洲鼓" (*fei1 zhou1 gu3*, "African drum"). The CoT rationale encounters two issues: incorrectly dividing the multiword term into two parts, "非珠" and "牛" instead of the correct pattern "非" and "珠牛", and excessively restoring the character "牛". In contrast, MoT prompts stress the entire word, facilitating correct restoration.

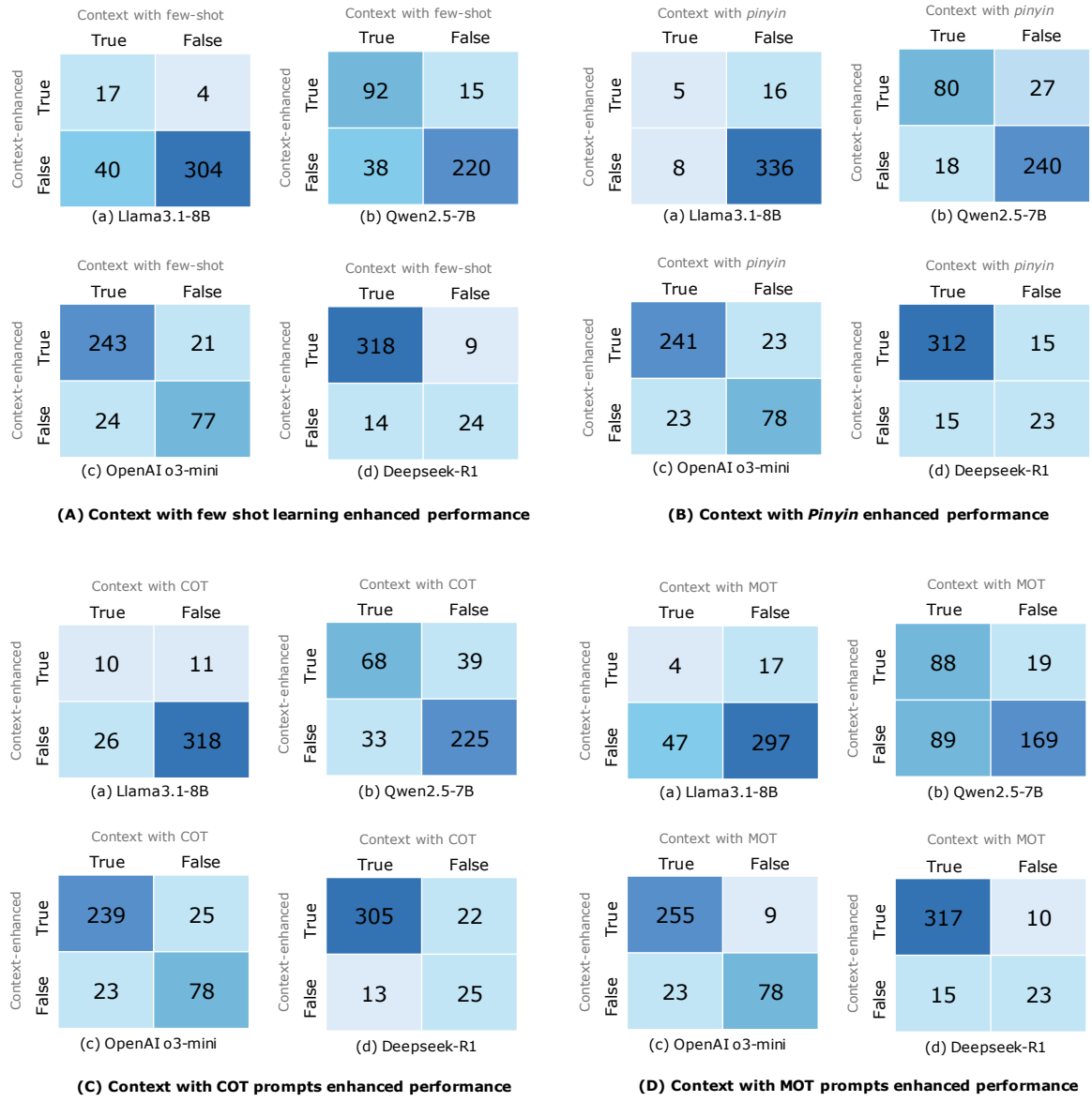

Figure 9: Comparison context-enhanced prompt between context enhanced with few-shot learning prompt, context enhanced with *pinyin* syllable prompt, context enhanced prompt with COT strategy, and context enhanced prompt with MOT experiments' results, in the form of confusion matrices.

Type	Homophone	CoT Rationale
Same semantic meaning restored word	石乐志 (shi2 le4 zhi4) (stone happy ambition)	{“推理过程”：“句子中的‘石乐志’发音与’失智’（shi1zhi4）相近，‘失智’指失去理智或精神恍惚，符合后文‘忘东忘西’的健忘语境。网络用语中常用‘石乐志’谐音表达这种状态。”,“原词”：“失智”}
Similar meaning with elements lost restored word	雾化女性 (wu4 hua4 nü3 xing4) (atomization female)	{“推理过程”：“原句中的‘雾化’与‘物化’发音相同，‘物化’指将人视为物品，常用于批评性别歧视行为，符合句意中对男性行为的负面评价。而‘雾化’的字面含义不符合语境，因此判断为谐音替换。”,“原词”：“物化”}
Complete wrong restoration	非珠牛 (fei1 zhu3 niu2) (Non jewelry cow)	{“推理过程”：“谐音词‘非珠牛’的发音为fei1 zhu3 niu1，需寻找与之发音相近且符合语境的音乐类型。‘非珠’与‘非洲’发音接近（zhu1与zhou1在某些方言中相似），‘牛’可能是‘鼓’的误写或方言发音(gu3与niu2可能存在地域性发音混淆)。结合‘非洲鼓音乐’在年轻人中流行的文化现象，判断原词应为‘非洲鼓’。”,“原词”：“非洲鼓”}
Type	Word	MoT Rationale
Same semantic meaning restored word	石乐志 (shi2 le4 zhi4) (stone happy ambition)	{“推理过程”：“‘石乐志’的拼音是shi2 le4 zhi4，原词应该为‘失了智’，拼音是shi1le5 zhi4。这是属于部分同音字置换形成的谐音词现象，其中‘石’与‘失’存在声调差异（阳平变阴平），‘乐’与‘了’属于同音字置换（仅改变汉字写法），‘志’与‘智’属于完全同音字置换。这种谐音现象通过调整部分字的声调（shi2→shi1）并配合同音字替换（le→le，zhi→zhi）来实现谐音效果，在保持整体发音近似性的同时形成网络用语的特殊表达方式。”,‘原词’：“失了智”}
Similar meaning with elements lost restored word	雾化女性 (wu4 hua4 nü3 xing4) (atomization female)	{“推理过程”：“‘雾化’的拼音wu4 hua4，原词应该为‘物化’，拼音为wu4 hua4。这是属于完全的同音字置换形成的谐音词现象，拼音拼写(发声位置)以及音调没有发生任何变化。这个谐音词中的‘雾’属于遭到置换的字。将‘物’替换为‘雾’。这一现象仅改变了汉字写法，保持发音一致形成了谐音效果”,“原词”：“物化”}
Complete wrong restoration	非珠牛 (fei1 zhu3 niu2) (Non jewelry cow)	{“推理过程”：“‘非珠牛’的拼音是fei1 zhu1 niu2，原词应该为“非主流”，拼音是fei1 zhu3 niu2。这是属于发音位置变化结合声调变化的谐音词现象。其中‘珠’字将原词‘主’(zhu3)的声调从第三声改为第一声，‘牛’字则将‘流’(liu2)的声母从[l]改为[n]，在部分方言区存在n/l不分现象。这一谐音词通过调整声母发音位置(n/l替换)和改变声调(第三声转第一声)，在保持整体音节相似度的同时形成了诙谐表达效果”,“原词”：“非主流”}

Table 11: This table shows the three types of words and examples, which represent three kinds of restoration error. The Chinese original CoT and MoT prompts are listed in the right part of the table with the rationale procedure and restored words.

Type	Homophone	CoT Rationale
Same semantic meaning restored word	石乐志 (shi2 le4 zhi4) (stone happy ambition)	Reasoning process: The term "石乐志" in the sentence has a similar pronunciation to "失智", shi1 zhi4. "失智" means losing one's rationality or being in a trance, which is consistent with the context of forgetfulness described as "forgetting things here and there" later. In Internet lang, "石乐志" is often used as a homophonic way to express this state. Original term: "失智" (loss of wisdom) (shi1 zhi4)
Similar meaning with elements lost restored word	雾化女性 (wu4 hua4 nü3 xing4) (atomization female)	Reasoning process: In the original sentence, the pronunciation of "雾化" (atomization) is the same as that of "物化", wu4 hua4. It aligns with the negative evaluation of men's behavior in the meaning of the sentence. However, the literal meaning of "雾化" does not fit the context. Therefore, it is determined that it is a homophonic substitution. Restored word: "物化"(objectify) (wu4 hua4)
Complete wrong restoration	非珠牛 (fei1 zhu3 niu2) (Non jewelry cow)	Reasoning process: The homophonic word "非珠牛" is pronounced as fei1 zhu1 niu2. We need to find a musical genre with a similar pronunciation that fits the context. "非珠" has a pronunciation close to fei1 zhu1 (the pronunciation of zhu1 and zhou1 can be similar in some dialects), and niu2 might be a miswriting of or a dialectal pronunciation of "鼓" (there could be a confusion in regional pronunciations between gu3 and niu2). Considering the cultural phenomenon that African drum music is popular among young people, it is judged that the original word should be "非洲鼓" (African drum), fei1 zhou1 gu3. Restored word: "非洲鼓" (African drum) (fei1 zhou1 gu3)
Type	**Word**	**MoT Rationale**
Same semantic meaning restored word	石乐志 (shi2 le4 zhi4) (stone happy ambition)	Reasoning process: The *pinyin* of "石乐志" is shi2 le4 zhi4, and the original phrase should be "失了智" with the *pinyin* shi1 le5 zhi4. This is a phenomenon of homophonic words formed by replacing some homophonic characters. Among them, there is a difference in tone between "石" (shi2) and "失" (shi1); "乐" and "了" are homophonic character replacements(only the written Chinese characters are changed); "志" and "智" are completely homophonic character replacements. This kind of homophonic phenomenon is achieved by adjusting the tones of some characters (shi2→shi1) and matching with homophonic character replacements (le→le, zhi→zhi). While maintaining the approximate similarity of the overall pronunciation, it forms a special expression in Internet slang. Restored word: "失了智" (loss of wisdom)(shi1 le5 zhi4)
Similar meaning with elements lost restored word	雾化女性 (wu4 hua4 nü3 xing4) (atomization female)	Reasoning process: The *pinyin* of "雾化" (atomization) is wu4 hua4, and the original word should be "物化" (objectify), with the *pinyin* also being wu4 hua4. This is a phenomenon of homophonic words formed by a complete replacement of homophonic characters, where neither the *pinyin* syllables spelling (place of articulation) nor the tone has changed at all. In this homophonic word, the character "雾" (wu4) is the one that has been replaced. The character "物" (wu4) has been substituted with "雾" (wu4). This phenomenon only changes the written form of the Chinese character while keeping the pronunciation consistent, thus creating a homophonic effect. Original word: "物化" (objectify) (wu4 hua4)
Complete wrong restoration	非珠牛 (fei1 zhu3 niu2) (Non jewelry cow)	Reasoning process: The *pinyin* of "非珠牛" is fei1 zhu1 niu2, and the original word should be "非主流" with the *pinyin* fei1 zhu3 liu2. This is a phenomenon of homophonic words that combines changes in the place of pronunciation and tone changes. Among them, for the character "珠" (zhu1), the tone of the original character "主" (zhu3) has been changed from the third tone to the first tone. As for the character "牛" (niu2), the initial consonant of "流" (liu2) has been changed from [l] to [n]. There is a phenomenon of confusion between "n" and "l" in some dialect areas. This homophonic word forms a humorous expression effect while maintaining the overall similarity of syllables by adjusting the pronunciation position of the initial consonant (replacement of "n" and "l") and changing the tone (changing from the third tone to the first tone). Original word: "非主流" (non-mainstream)(fei1 zhu3 liu2)

Table 12: This table shows the three types of words, which represent three kinds of restoration error. The English-translated CoT and MoT prompts are listed in the right part of the table with the rationale procedure and restored words.