Radical Prompting: Enhancing Chinese Language Models with Character Visual Analysis

Anonymous ACL submission

Abstract

As a glyphic language, Chinese incorporates information-rich visual features, with distinct characters combining to form compounds that inherit the meaning or pronunciation of their components. However, we argue that Large Language Models (LLMs) fail to effectively harness this valuable feature. This study designs 'radical prompting' to improve LLMs' effectiveness across general NLP tasks such as Part-Of-Speech (POS) tagging and investigates the limitations of contemporary LLMs in accurately identifying the visual information of characters. Results demonstrate that the introduction of 'radical prompting' markedly improved LLM performance across various NLP tasks, particularly when correct radicals were provided, highlighting its potential as a crucial tool for optimizing Chinese language processing. However, most LLMs struggle to correctly identify the visual fundamentals of Chinese characters, which limits their effectiveness. Despite some progress achieved through prompting and fine-tuning, the current accuracy levels still fall short of the desired excellence.

1 Introduction

Unlike alphabetical languages, a character is not the smallest meaningful unit in Chinese. Most Chinese characters consist of meaningful radicals or components, which can themselves be made up of smaller radicals and characters. For example, the Chinese character "花" (meaning "flower") is composed of the "艹" (grass) radical, which contributes to its semantic property, and the component "化," which guides its pronunciation. The component "化" can be further decomposed into the "亻" (human) radical and the "七" component as illustrated in Figure 1. Therefore, when encountering unknown or unfamiliar characters, it is a very common and useful strategy to look at the radicals to estimate their meanings or pronunciations. Often, the components within unknown characters are simpler and more familiar. Inspired by this strategy, we designed the radical prompting approach to implement a similar strategy in Chinese NLP tasks in Section 4. For part-of-speech (POS) tagging task, this method effectively improves LLMs' performance. For instance, when providing a gold label, GPT-3.5 Turbo experiences approximately a 14% improvement. For more challenging tasks like named entity recognition (NER), larger and more robust models, such as Claude-3, achieve around a 6% improvement even without being provided the correct radicals. Smaller models, however, show a minor increase or even a decrease in performance.

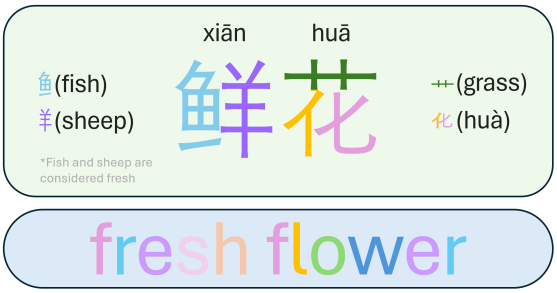

Figure 1: component of "fresh flower" in Chinese and English

To further investigate the improvement gap between providing the actual radical and not providing it, we initiated an intrinsic evaluation of LLMs' ability to identify the visual information of Chinese characters in Section 5. We constructed a dataset containing Chinese characters and three information-rich properties embedded within them: the components or radicals of the characters, the structure composing the characters, and the total stroke count of the characters. The components serve as foundational elements, akin to prefixes or suffixes in alphabetical languages, and provide clues to both the meaning and pronunciation of the characters. The structural composition of the characters, categorized into eight distinct types as

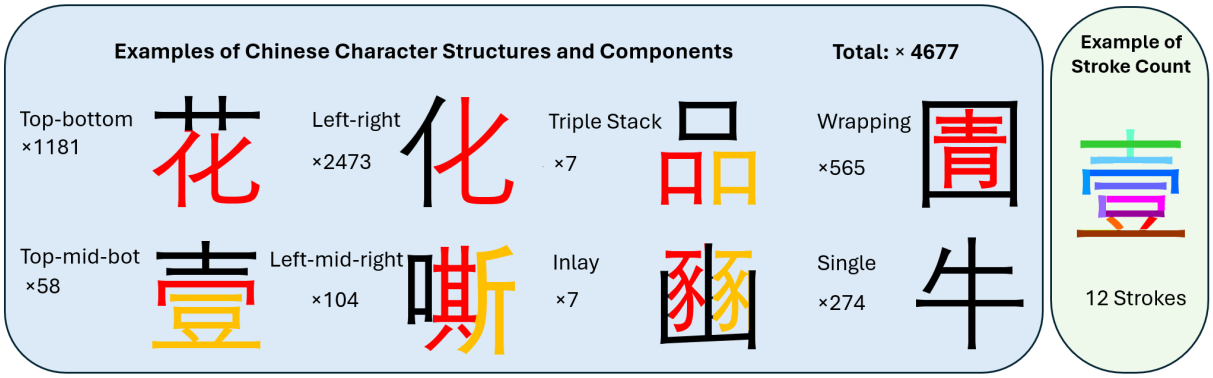

Figure 2: The statistics and examples of our Chinese character visuals dataset. While the total number of Chinese characters in existence far exceeds the scope of our dataset, it's worth noting that the List of Commonly Used Characters in Modern Chinese comprises only 1,000 primary characters and an additional 2,500 less frequently used ones. Our dataset goes beyond the number of conventional set.

shown in Figure 2, influences how characters are perceived, specifically affecting the order in which a character's components are recognized. Lastly, the stroke count offers a measure of a character's visual complexity or density. Unlike alphabetic languages, where word length can hint at complexity, Chinese characters occupy uniform space, making stroke count a valuable indicator of intricacy between complex and simple characters. The results show that all LLMs, whether Chinese or multilingual, failed to successfully identify this visual information, resulting in relatively high entropy and low F1 scores, which indicate low confidence and poor performance on the task. Traditionally, most LLMs process Chinese text at various levels: at the character level using Unicode, at the word level, or at an intermediary level through techniques like Byte Pair Encoding (BPE). This processing approach tends to filter out the explicit visual information inherent to the components of characters, which need to be captured at a more fine-grained level to enhance understanding. In response to these challenges, we explored various architectures, fine-tuning methods, prompting strategies, and encoding methods. Our investigation revealed that pixel-level encoders and glyph-based encoding are particularly effective in handling this task, suggesting directions for further research.

This paper makes four key contributions to underscore the importance of the issue and suggest avenues for future research: 1) It develops a dataset that captures the visual aspects of Chinese characters; 2) introduces 'radical prompting' to enhance the performance of LLMs across various NLP tasks; 3) examines the challenges contemporary LLMs face in precisely recognizing the visual information of characters; and 4) explores novel methods to boost contemporary LLMs' capabilities in identify visual structure of the characters.

2 Related Work

Within the realm of Chinese language processing, the exploration of radicals has been relatively unexplored. Therefore, our discussion in the related work section will focus on research that intersects with the study of radicals, spanning topics from Chinese character decomposition in computer vision, to specialized datasets, and existing strategies that use radicals in language processing.

Chinese Character Decomposition in Computer Vision The task of decomposing Chinese characters into their constituent components closely aligns with challenges faced in the field of computer vision. Research within this domain, such as the studies by (Ma et al., 2021), (Xia, 1994), (Liu et al., 2021), has explored analogous challenges. The work by (Zhang et al., 2018) employs a methodical approach by categorizing characters into structured types, such as top-bottom or left-right, and further decomposing sub-components according to their spatial arrangements—akin to the layered structural analysis this paper adopts. This technique allows for a nuanced breakdown of characters into constituent elements, as will be further explored in Section 5.1.

Chinese decomposition dataset In reviewing available resources, we encountered a comprehensive dataset (CJKVI) that offers decompositions for the unified characters of Chinese, Japanese,

and Korean. Although this collection encompasses enormous Chinese characters, it does not cite any authoritative sources for its data. This omission leads to potential ambiguity due to multiple decomposition sequences for individual characters.

Contrastingly, our approach utilizes the structural classifications from the Kangxi Dictionary, ensuring a validated framework for segmentation and maintaining a manageable dataset size to guarantee accuracy. Additionally, our dataset stands apart by gathering stroke count data from the Xinhua Dictionary, thus creating a dataset specifically focus on visual information of Chinese characters.

Glyphic Embedding Strategies in LMs Recent studies have increasingly sought to leverage the rich visual information inherent in Chinese characters to enhance language model performance. For instance, (Sun et al., 2021) introduce a novel approach that incorporates different embeddings alongside glyph embeddings derived from different fonts to enrich character representations. Similarly, Si et al. (2021) delve into the potential of stroke encoding among other glyph based input method to explore their performance (Si et al., 2021). Additionally, (Shi et al., 2015) harness radical information, utilizing it as a key component for embedding Chinese characters. These systems share a common challenge: the necessity of retraining the entire system which not only demands substantial computational resources but also raises questions about scalability and adaptability, especially since these enhancements have predominantly been applied to smaller-scale models. Our paper, in contrast, zeroes in on the impact of incorporating visual features of Chinese characters, such as stroke count and structure, directly within contemporary large language models, bypassing the complex embedding strategies employed by earlier studies.

3 Dataset

To evaluate the proficiency of contemporary language models with character fundamentals, we curated a dataset from simplified Chinese characters sourced from the digitized Kangxi Dictionary.[1] These characters were categorized into **8** distinct structures, as defined by the dictionary, and are detailed in Figure 2. With the assistance of APISpace's Chinese character segmentation API and the Xinhua Dictionary API, we identified the components and stroke count for each character. The results underwent manual verification to ensure the dataset's accuracy and integrity. In the segmentation process, we not only adhered to the established practice of segmenting by structures, as detailed in (Zhang et al., 2018), but also made a concerted effort to retain meaningful units wherever feasible. For example, "人" could technically be identified as a left and right structure; however, doing so would reduce it to meaningless strokes. Therefore, we classify "人" as a single structure and do not separate it further.

[1]The Kangxi Dictionary has been regarded from its inception until the early 20th century as the preeminent reference for written Chinese characters. It has undergone updates since its original publication, maintaining its esteemed position in the realm of Chinese lexicography.

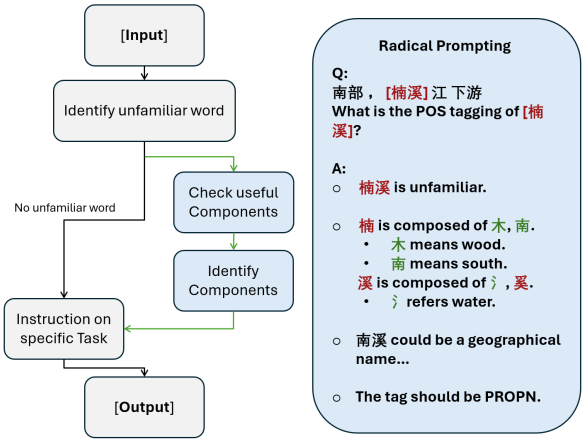

Figure 3: The process of radical prompting and example of radial prompting answer for part-of-speech (POS) tagging with an unfamiliar Chinese word.

4 Extrinsic Evaluation with Radical Prompting

In this section, we extrinsically examine the significance of radicals, the key visual feature in Chinese characters, by prompting models to leverage radical knowledge as illustrated in Figure 3. This approach aims to assess the impact of such visual properties on improving Chinese language processing tasks.

4.1 Tasks

Part-of-Speech (POS) tagging. In evaluating the effectiveness of Large Language Models (LLMs) in Chinese Part-Of-Speech (POS) tagging, three datasets were utilized: the GSD Simplified dataset for contemporary Chinese (Qi and Yasuoka, 2023), the Parallel Universal Dependencies (PUD) dataset (McDonald et al., 2023) for comparative modern language analysis, and a novel dataset derived from

Model	POS Tagging GSD			NER People's Daily		CWS GSD	
	Baseline	RP	RP (Oracle)	Baseline	RP	Baseline	RP
GPT-3.5	59.08	64.62(+5.5)	67.56(+8.5)	56.89	55.97(-0.9)	95.68	94.87(-0.8)
GPT-4	71.55	72.14(+0.6)	72.95(+1.4)	66.04	68.05(+2.0)	94.21	94.88(+0.7)
Claude-3	69.37	70.68(+1.3)	70.45(+1.1)	69.74	73.79(+4.1)	94.90	95.16(+0.3)
QWen 72B	62.20	65.38(+3.2)	67.32(+5.1)	62.73	59.59(-3.1)	96.59	95.57(-1.0)
ERNIE-Lite	27.06	24.97(-2.1)	32.73(+5.7)	12.10	12.99(+0.9)	88.04	88.70(+0.3)
Aya	68.86	68.91(+0.1)	70.41(+1.6)	38.24	36.36(-1.9)	87.98	89.08(+1.1)

Table 1: Comparison of model performances across various NLP tasks with baseline, radical prompting without golden components (RP), and radical prompting with oracle information (RP (Oracle)).

500 sentences in poems form the Tang Dynasty[2], annotated using Classical Chinese RoBERTa (Yasuoka, 2023). For this task, a 5-word span from each sentence was selected and the model was tasked with predicting the tag for the central word. We designed two versions of the task: one that supplies the correct component and radical information of the central word, and another that prompts the model to utilize radical information without explicitly providing it. We use F1 score to measure models' performance on this task.

Named Entity Recognition (NER). In assessing NER capabilities, this study examines performance across two distinct datasets: the People's Daily dataset (Chen, 2023), which focuses on formal Chinese text, and the Weibo NER dataset (Peng and Dredze, 2015), which is oriented towards casual and online Chinese text. Both datasets include tags for PER (person), LOC (location), and ORG (organization), with the Weibo NER dataset additionally incorporating GPE (Geo-Political Entities). While the Weibo NER dataset extends to annotate nominal entities, this task concentrates solely on traditional named entities. Both datasets adhere to the BIO tagging standard, facilitating a consistent evaluation framework. Similar to POS tagging, this task will be evaluated using the F1 score. Unlike POS tagging, where the focus is on a central word, NER involves labeling any word in a sentence, making it impractical to provide radical information for each word and character. Therefore, we will only evaluate the efficacy of radical prompting without supplying the correct component information.

[2]The Tang Dynasty is a period known for its well-preserved and flourishing poetry. The choice of classical poems is motivated by the precision and compactness of information in each character typical of this era, suggesting that more information is preserved at the subcharacter level—namely, in the radicals.

Chinese Word Segmentation (CWS). CWS is a unique task in Chinese language processing. Distinguished from many other languages, Chinese does not use delimiters such as spaces to separate words within sentences. Accurately segmenting text into individual words is critical, particularly for enhancing performance in further language processing tasks such as information extraction and machine translation (Peng et al., 2004). In this study, we utilize the same datasets as employed for the POS tagging task: the GSD and PUD from the Universal Dependencies collection. As the Universal Dependency datasets already separate sentences into words to tag their pos tag and dependencies. In this task, we give the whole sentence and ask the model to separate the sentence by words. To assess the effectiveness of models in this critical task, performance is evaluated using the F1 score.

4.2 Method

In our exploration of enhancing Chinese language processing through visual cues, we introduce a novel prompting method termed "radical prompting." This technique builds upon the foundation of the chain of thought (COT) prompting framework, which guides models through tasks in a sequential, step-by-step manner. The process begins with the model identifying any unclear words within a given context. Following this initial step, the model is instructed to dissect these words into their constituent components, specifically focusing on radicals. It then evaluates whether these components impart additional, useful information that can aid in task completion. Subsequently, the prompt guides the model to execute specific tasks, attempting to gain from radical analysis to enhance task performance.

A crucial aspect of radical prompting is the emphasis on cautious and judicious use of compo-

nent information. This cautionary note is vital due to the historical evolution of Chinese characters, where some have acquired meanings that diverge significantly from their original components. Therefore, models are advised to critically assess the relevance and accuracy of the information provided by character components, especially in cases where the linkage between form and meaning may not be straightforward. Additionally, the model is prompted with one example where the information from radicals may not be pertinent, highlighting the importance of discerning the applicability of radical information.

A description of the radical prompting lines utilized for the Part-of-Speech (POS) tagging task is available in the appendix, see Section B.2.

4.3 Experiment Setup

In our exploration of the impact of radical prompting on Chinese language processing, we carefully selected a suite of models for evaluation: GPT-3.5, GPT-4, Claude-3, QWen-1.5 72B Chat, Aya, and ERNIE-Lite-8K. To assess their performance, we applied these models to the specific NLP tasks described in Section 4.2, each associated with its own task-specific dataset. The experiment involved processing a 1,000 sentences from these datasets 5 times for consistency and reliability of the results across all models, with the exception of GPT-4. Due to the higher operational costs associated with the GPT-4 and Claude-3 API, the experiment was adjusted to include only 500 sentences for those two models.

4.4 Result and Error Analysis

We observed a general enhancement in model performance across all tasks shown in Table 1. This improvement was particularly notable in POS tagging tasks, where the provision of the correct component for the central word further augmented performance: GPT-3.5 Turbo experience a 14% improvement when implemented with radical prompting with correct component provided.

A detailed error analysis sheds light on the nuances behind the observed improvements in POS tagging tasks as illustrated in Appendix B.1. Our data indicates that for GPT-3.5-Turbo, whether radical prompting was applied, the number of cases correctly identified where components were not examined remained relatively stable. However, when the model utilize radical information by identifying central word as unfamiliar, an additional 81.2 cases were correctly identified compared to the baseline. The advanced models GPT-4 and Claude-3, while showing improvement, did not exhibit a large margin of change. This can be attributed to their less frequent detection of unfamiliar words, lowering tendency to leverage radical information.

For NER and CWS tasks, the impact of radical prompting exhibited a nuanced relationship with model size and capacity. Smaller models, such as GPT-3.5, experienced a slight decline in performance following the introduction of radical prompting. Conversely, more robust models like GPT-4 and Claude-3 demonstrated marked improvements in their NER and CWS task outcomes. One plausible explanation for this trend is the prevalence of transliterated foreign terms in Chinese—which are adapted based on pronunciation rather than meaning. Since radical information offers little to no help in deciphering these terms, their frequent occurrence as entities in the text might confound the model, negating the advantages of radical prompting in these instances. Another contributing factor to the nuanced performance is the inherent complexity of the NER and CWS tasks. These tasks require the precise identification of a wide array of components to accurately determine the meanings of words in their specific contexts. Given the diversity of Chinese characters and the nuances of context, supplying the model with complete and accurate component information is a formidable task. It is reasonable to hypothesize that, as seen with POS tagging, if oracle information about components could be provided, these tasks might exhibit improved performance as well.

5 Intrinsic Evaluation on Chinese Character Visuals

Build on previous findings, this section aims to delves into the performance of LLMs on the visual recognition task such as components recognition. We begin with a baseline assessment of these models' performance on predefined tasks, focusing on their ability to process and interpret the complex visual structure of Chinese characters. Following this, we examine the efficacy of various enhancement techniques. The section will detail the tasks evaluated, describe the two experimental approaches undertaken, and culminate in a comprehensive analysis to distill further insights into the effectiveness of the visual enhancement techniques applied.

Model	Structure		Component						
	F1	Entropy	1st Pos		2nd Pos		3rd Pos		Overall F1
			Acc	Entropy	Acc	Entropy	Acc	Entropy	
GPT-3.5 Few	19.71	0.84	28.55	0.52	17.74	1.16	2.45	0.85	45.29
GPT-3.5 Zero	22.82	0.88	33.61	0.79	19.09	1.40	5.48	0.69	48.86
GPT-4 Few	45.28	0.48	58.22	0.14	31.67	0.52	20.48	0.22	68.57
GPT-4 Zero	35.40	0.54	58.66	0.22	31.03	0.89	11.50	0.34	67.82
ERNIE-Lite	38.63	0.42	33.94	0.71	9.41	1.35	0.00	1.29	34.90
Yi-6B	27.42	1.17	47.31	1.72	19.19	1.56	0.00	0.80	34.21
Qwen-7B	29.31	1.28	33.42	1.82	21.34	1.51	3.42	1.50	20.13
Baichuan-13B	24.12	0.81	38.41	1.45	24.49	1.53	4.21	0.57	24.17
Mistral-7B	27.77	1.64	42.11	2.20	27.95	1.44	2.49	0.37	28.67

Table 2: LLMs' Performance in Structure and Component recognition of Chinese characters.

Model	Stroke Count	
	MSE	MAE
GPT-3.5-Turbo Few	23.18	1.79
GPT-3.5-Turbo Zero	52.89	2.06
GPT-4 Few	23.18	1.79
GPT-4 Zero	12.17	1.99
Ernie-Lite	33.49	4.49
Yi-6B	29.49	4.24
Qwen-7B	34.16	4.62
Baichuan-13B	32.70	4.31
Mistral-7B	165.39	10.94

Table 3: LLMs' Performance in Stroke Count Identification.

5.1 Tasks

Structures recognition of Chinese character. We assess LLMs' ability to identify the correct structural arrangements of Chinese characters in our dataset, with performance evaluated using the F1 score. We categorize all Chinese characters into eight major structural arrangements: top-bottom, left-right, top-mid-bottom, left-mid-right, wrapping[3], inlay, triple-stack, and single structure (those that cannot be further segmented), as detailed in Table 1. The structure of Chinese characters can be complex, with multiple layers of structure compounding upon each other. For example, the character "花," as illustrated in Figure 1, primarily presents a top-bottom structure, segmented into "艹" and "化." Upon closer inspection, "化," which exhibits a left-right structure, can be further decomposed into "亻" and "七." To maintain consistency in our segmentation approach, we segment characters based on their primary structure, ensuring uniformity across our dataset.

Components recognition of Chinese character. We evaluate LLMs' proficiency in accurately identifying the components of Chinese characters within our dataset. In this particular task, models are not mandated to explicitly determine the structural arrangement of characters. Nonetheless, they are expected to recognize and output the character's components in their correct order of perception, which inherently relates to the character's structural arrangement. Similar to the structure task, we apply a uniform segmentation principle that focuses on identifying characters' primary structures without delving into further sub-component breakdown. For instance, the character "花" exhibits a top-bottom structure, with "艹" positioned at the top and "化" at the bottom. Accordingly, our segmentation isolates these two principal components only, avoiding decomposition beyond this primary structural division. To evaluate performance comprehensively, positional accuracy for the first, second, and third components is assessed, with the overall F1 score calculated focusing on correct predictions, regardless of their positional order.

Stroke count identification of Chinese character. We evaluate the LLMs' proficiency in accurately determining the stroke count of Chinese characters from our dataset. Here, the models are tasked with producing a single integer value representing the total number of strokes required to write each char-

[3]While the category of wrapping structures can divide further, for clarity and due to their similar order of visual perception, we have amalgamated all types of wrapping into one comprehensive 'wrapping' structure.

acter. The accurate assessment of stroke count is critical as it provides a measure of the character's complexity. To quantitatively measure the models' performance on this task, we will use the Mean Absolute Error (MAE) as our metric.

5.2 Evaluating Standard LLMs on Chinese Character Visual Recognition

Setup. To explore LLMs' abilities in recognizing the visual and structural intricacies of Chinese characters, 2,000 Chinese characters were randomly selected from our dataset to serve as query characters for three distinct tasks. This selection was strategically repeated five times to ensure the reliability in the performance assessment, focusing on models' consistency and confidence through entropy calculation of the results. Importantly, characters with a 'Single Structure' were omitted from the component recognition task due to their inherent segmentation constraints.

The models selected for this evaluation encompass a diverse range, including Yi-6B, Qwen-7B-Chat, Baichuan-13B, Mistral-7B, ERNIE-Lite, GPT-4, and GPT-3.5 Turbo. To thoroughly examine the GPT models' understanding, they were subjected to zero-shot, and few-shot testing scenarios. Specifically, in the few-shot setting, models were given a representative example for each of the eight structures. For the remaining models, the evaluation was only placed on the few-shot tests due to the challenge of task completion without example guidance.

Results. Our examination of model performance on the visual complexities of Chinese characters, as shown in Tables 2 and 3, reveals GPT-4 as the most capable model among those tested, yet its performance still falls short of being considered optimal. Specifically, GPT-4's structure identification F1 score, which stands at 45.28, demonstrates significant challenges in accurately discerning character structures. Additionally, for component recognition, the overall F1 and accuracy for each position is suboptimal, with entropy values indicating varying levels of confidence across predictions. This decline is notably pronounced when moving from the first to the third component, reflecting the model's difficulties in identifying all components of characters accurately. In stroke count estimation, GPT-4 demonstrates a Mean Squared Error (MSE) of 12.17 and a Mean Absolute Error (MAE) of 1.99. Given that the average stroke count of characters in our dataset is around 10, these error rates underscore the models' imprecision in capturing the exact stroke count of characters.

Upon delving into the results across different structures, as detailed in Table 8, we observe a specific challenge: models struggle to differentiate between structures with three components and their two-component counterparts, such as distinguishing top-bottom from top-mid-bottom arrangements.

Viewing the visual tasks collectively, our analysis revealed a significant pattern: models that demonstrate proficiency in one visual aspect—be it component recognition, structure identification, or stroke count estimation—tend to exhibit improved performance in the other areas as well. This suggests a shared underlying ability among models to process visual information effectively.

5.3 Advanced Techniques in Chinese Character Structure Recognition

Setup. This experiment focuses on evaluating the efficacy of various methods, such as visual-related architectures, fine-tuning, and prompting, on the Chinese structure recognition task.

We investigate two distinct architectural approaches aimed at enhancing task performance: vision-integrated multi-modal models and pixel-based encoder language models. The first approach is represented by GPT-4 Vision and Claude-3 Vision , where we incorporate character images as part of the input, assessing its performance on a dataset of 2,000 randomly selected characters. The second approach is embodied by the PIXEL model, a pixel-based encoder language model ((Rust et al., 2023)), distinctively not a large language model but rather a focused language model. This model is trained exclusively on the English Wikipedia corpus and undergoes specific fine-tuning and testing for structure recognition within a span-based question-answering (QA) framework, utilizing 70% of our dataset for training and rest for evaluation.

Furthermore, we delve into the efficacy of prompting on GPT-3.5 to capture the visual specificity of Chinese characters. The prompting strategy involves guiding GPT-3.5 to identify the radical of a character—leveraging the association between radicals and character meanings—before prompting it to outline the character's other components. This method is assessed using a separate set of 2,000 randomly selected characters. Detailed descriptions of the specific prompting lines employed are available in the appendix 4. Finally, GPT-3.5

Model	Structure F1
GPT-3.5(Zero)	19.71
GPT-3.5 Fine-tuned	64.76
GPT-3.5 Structure Prompting	38.08
GPT-4 Vision	37.02
Claude-3 Vision	26.09
PIXEL Fine-tuned	**84.57**

Table 4: Performance Improvement on Different Method on Structure Recognition Task

Encoding	F1 score
Unicode	39.80
Stroke	**43.80**
PinYin	13.85
WuBi	11.81
CangJie	11.66

Table 5: GPT-3.5 Fine-tuning' Performance on different way of encoding.

undergoes fine-tuning to assess the enhancement of its performance on visually specific tasks.

Results. The comprehensive evaluation of models on the Chinese character structure recognition task is encapsulated in Table 4:

A noteworthy aspect of our comparison involves the performance of models incorporating visual information processing capabilities. An interesting observation arises from the parallel efficiency observed between GPT-4 Vision and GPT-4 in the few-shot scenario. The similarity in their performance might hint at the vision component of the architecture not being specifically trained on tasks related to Chinese character structure recognition, which could limit its effectiveness in leveraging visual data for this purpose. On the other hand, the PIXEL model achieves an exceptional F1 score of 84 after fine-tuning.

The employment of structured reasoning prompts, which directed the model to engage in a more thorough analysis of character structures, resulted in a notable performance uptick. Specifically, this strategic refinement elevated the model's F1 score to approximately 38. Moreover, fine-tuning GPT-3.5 with a dataset specifically curated for this investigation significantly advanced the model's proficiency in recognizing Chinese characters. After fine-tuning, the model demonstrated a remarkable F1 score of 62, showcasing its potential to adapt and master the visual intricacies of Chinese characters. Nonetheless, despite these substantial improvements, the model's performance fell short of reaching the F1 score benchmark of 70.

5.4 Encoding Analysis

Building on the positive outcomes of fine-tuning GPT-3.5, we extended our research to delve into the model's capacity for learning by examining the effects of utilizing different encoding.

Setup. We fine-tuned GPT-3.5 while explicitly switching all Chinese characters to various encoding—namely, Unicode, stroke, pinyin[4], Wubi, and Cangjie[5]—to evaluate the extent to which these representations impact the model's proficiency in internalizing visual knowledge of Chinese characters. results are shown in Table 5.

Results. The results indicated that Unicode encoding perform comparably to stroke encoding, which are rich in glyphic information, and significantly outperform pinyin encoding, which are limited to phonetic information. This disparity suggests that Unicode, despite its abstract nature, carries implicit visual cues. The structured arrangement of Chinese characters in Unicode, predicated on the stroke number of the radical and subsequent components, mirrors the visual characteristics intrinsic to these characters as shown in Table 9. However, the full potential of Unicode is somewhat diminished by a multitude of exceptions and a broad spectrum of extensions that complicate its utility in conveying structured visual knowledge.

6 Conclusion

In this paper, we commenced our exploration with an in-depth examination of radical prompting and its impact on enhancing the performance of Large Language Models (LLMs) in general NLP tasks. This led us to evaluate the ability of LLMs to recognize radicals within Chinese characters, where we observed their suboptimal performance, underscoring critical importance of the visual aspects of Chinese characters in advancing the processing capabilities of LLMs for Chinese language tasks. We finish our paper with point toward several promising directions for future research.

[4]Pinyin is the Romanization of the Chinese characters based on their pronunciation. In Mandarin, it's the standard method for typing Chinese characters.

[5]Wubi and Cangjie are two glyph based input method that are uncommon to use.

Limitations

Our study, while contributing valuable insights into the integration of radical prompting for Chinese language models, encounters several limitations that suggest directions for future research. First, the dataset employed does not encompass the full array of Chinese characters but is confined to commonly used characters. This selective coverage might affect the scalability of our findings to all Chinese characters [especially when greater model meets unknown or unfamiliar character, there is a chance that our dataset does not cover that character]. Additionally, the study primarily evaluates the effectiveness of radical prompting on a narrow selection of models and specific NLP tasks, which might not reflect its utility across different models or broader language processing applications.

Furthermore, an intrinsic limitation of our methodology arises from the exclusive use of English in our prompting lines. Incorporating Chinese in the prompting strategy could potentially enhance the relevance and effectiveness of prompts, aligning better with the linguistic context of the target language.

References

Han Chen. 2023. People's daily (renmin daily) named entity recognition dataset. http://paper.people.com.cn/. A comprehensive dataset from the People's Daily, covering news from 2021/01/01 to 2023/12/05, for Named Entity Recognition with news segments labeled for LOC, ORG, PER entities using BIO tagging strategy. License: CC0: Public Domain.

CJKVI. Cjkvi-ids: Ideographic description sequences for cjk unified ideographs. https://github.com/cjkvi/cjkvi-ids. Accessed: 2024-4-4.

Xiaodong Liu, David Wisniewski, L. Vermeylen, Ana F. Palenciano, Wenjie Liu, and M. Brysbaert. 2021. The representations of chinese characters: Evidence from sublexical components. *Journal of Neuroscience*, 42(1):135.

Jiefeng Ma, Zirui Wang, and Jun Du. 2021. An open-source library of 2d-gmm-hmm based on kaldi toolkit and its application to handwritten chinese character recognition. *Lecture Notes in Computer Science*, 12888.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Tackstrom, Claudia Bedini, Nuria Bertomeu Castello, and Jungmee Lee. 2023. Parallel universal dependencies (pud) treebanks for multilingual parsing. Available for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies. Annotations provided by Google and converted to UD v2 guidelines by the UD community.

Fuchun Peng, Xiangji Huang, Dale Schuurmans, and Nick Cercone. 2004. Investigating the relationship between word segmentation performance and retrieval performance in chinese ir. In *Proceedings of the School of Computer Science*, University of Waterloo, 200 University Ave. West, Waterloo, Ontario, Canada, N2L 3G1.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the Human Language Technology Center of Excellence*, Baltimore, MD. Johns Hopkins University.

Peng Qi and Koichi Yasuoka. 2023. Simplified chinese universal dependencies version 2.13. Universal Dependencies (UD) Chinese GSDSimp treebank. Available from GitHub: UD_Chinese-GSDSimp.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels. In *The Eleventh International Conference on Learning Representations*.

Xinlei Shi, Junjie Zhai, Xudong Yang, Zehua Xie, and Chao Liu. 2015. Radical embedding: Delving deeper into chinese radicals. In *Proceedings of the Association for Computational Linguistics (ACL)*. Sogou Technology Inc., Beijing, China.

Chenglei Si, Zhengyan Zhang, Yingfa Chen, Fanchao Qi, Xiaozhi Wang, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Sub-character tokenization for chinese pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:634–649.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. In *Proceedings of the Association for Computational Linguistics (ACL)*. Shannon.AI; Zhejiang University; Key Lab of Intelligent Information Processing of Chinese Academy of Sciences.

Franck Xia. 1994. Knowledge-based sub-pattern segmentation: decompositions of chinese characters. *Proceedings of the International Conference on Image Processing*.

Koichi Yasuoka. 2023. Roberta model pre-trained on classical chinese texts. https://huggingface.co/KoichiYasuoka/roberta-classical-chinese-large-char. Derived from GuwenBERT-large with character-embeddings for traditional/simplified characters. Suitable for tasks like sentence-segmentation, POS-tagging, dependency-parsing.

Jianshu Zhang, Yixing Zhu, Jun Du, and Lirong Dai. 2018. Radical analysis network for zero-shot learning in printed chinese character recognition. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Hefei, Anhui, P.R. China. IEEE.

A General Experiment Details

Model Versions and Snapshots The experiments incorporated different versions of widely recognized models to evaluate their performance in processing Chinese characters. The specific snapshots used for each model are as follows:

- **GPT-3.5** and **GPT-4** were used with the snapshot dated *2023-11-06*.
- **Claude** model's evaluation utilized the *2024-02-29* snapshot.
- **Ernie-Lite-8K** was tested using the *2023-09-22* snapshot.

Temperature Settings

- **Aya**, **Yi-6B**, **Qwen-7B-Chat**, **Baichuan-13B**, and **Mistral-7B** were set at a lower temperature of *0.3* as recommended.
- For **other models** not specifically mentioned, a temperature setting of *0.7* was used.

B Detailed Radical Prompting Result

B.1 Quantitative Analysis on POS tagging Accuracy

We provide a case analysis for POS tagging in Table 6.

Category	Baseline	RP (Oracle)
Correct& Comp		+81.2
Correct without	608.6	611.2
Incorrect & Comp		+81.2
Incorrect without	391.4	265.8

Table 6: Quantitative analysis of GPT-3.5-Turbo's POS tagging accuracy on the number of correct and incorrect predictions with and without the examination of components using radical prompting compared to the baseline.

B.2 Full Result of Radical Prompting Experiment

We provide result of radical prompting on more dataset in Table 7.

Task

Output the structure of chinese character in abbreviation defined below:
the structure of the character must be one of the following:
sx(上下结构)
zy(左右结构)
szx(上中下结构)
zzy(左中右结构)
bw(包围结构)
xq(镶嵌结构)
dy(单一结构)
pin(品字结构)

Let's think step by step. First identify the radical of the character. The radical is usually associated with the property of the character. Then, based on the relative position of the radical and remaining component of the character, identify the structure of the character. Clearly state the structure (one of above) in the end of your answer.

Finalize your choice in JSON format, where the key must be "structure" and the value must be one of the abbreviations of structure above.

Character to analyze:

Figure 4: Prompt Line to Enhance Recognition of Chinese Character Structures

B.3 Prompting Example

We provide our prompting lines for POS tagging tasks in Figure 5.

C Details on Structure Recognition

C.1 Structure Recognition Across Structures

We provide detailed result for structure recognition across different structures in Table 8.

C.2 Enhancing Structure Recognition through Prompting Techniques

We present the prompt lines used to enhance the task of structure recognition in Figure 4.

D Unicode Example

A portion of the Unicode table is presented in 9 to demonstrate the visual character information embedded within Unicode.

E Responsible NLP Miscellanea

E.1 Intent usage

In response to potential inquiries regarding the scope and legitimacy of our experiments, it is important to clarify that all aspects of our research strictly adhere to the intended use cases of the Large Language Models (LLMs) and the NLP task datasets employed. Furthermore, our use of these models and datasets complies fully with the usage policies of the APIs for each model involved.

E.2 Computational Experiments Cost

In our research, we utilized vLLMs for evaluation on Yi 6B, Mistral 7B, Baichuan 13B, and Qwen

Task	Model	Baseline	Radical Prompting (No Gold)	Radical Prompting (Oracle)
POS Tagging PUD	GPT-3.5	62.61	69.90	73.46
	GPT-4	76.20	76.72	77.35
	Claude-3	69.37	70.45	70.68
	QWen 72B	62.20	65.38	67.32
	Ernie-Lite	30.35	30.29	41.29
	Aya	73.87	77.21	76.95
POS Tagging Poem	GPT-3.5	53.51	59.22	61.39
	GPT-4	66.94	67.11	67.57
	Claude-3	65.53	66.20	66.71
	QWen 72B	55.63	57.78	59.54
	Ernie-Lite	44.19	42.17	49.07
	Aya	65.53	66.19	66.71
NER Weibo	GPT-3.5	36.65	36.64	
	GPT-4	43.83	44.68	
	Claude-3	45.64	46.86	
	QWen 72B	31.78	35.83	
	Ernie-Lite	6.72	6.90	
	Aya	37.88	30.83	
CWS PUD	GPT-3.5	93.91	93.70	
	GPT-4	94.24	95.63	
	Claude-3	94.12	94.96	
	QWen 72B	89.79	91.94	
	Ernie-Lite	69.54	73.57	
	Aya	88.68	91.05	

Table 7: Comparison of model performances across various NLP tasks with baseline, radical prompting without golden components, and radical prompting with oracle information.

7B with a single a40 GPU. For other models, we accessed them through their respective APIs. The cost and running time for each model varied significantly. Specifically, the time required to run a single evaluation ranged from approximately 2 to 8 hours.

E.3 Avoid Data Leakage

For all NLP tasks assessed in this study, evaluations were exclusively conducted on the development sets of the respective datasets to prevent data leakage.

E.4 Personally Identifying Info

The dataset we created for evaluating the visual information of Chinese characters does not contain any offensive content or personally identifying information. However, we acknowledge the presence of individual names in the Weibo NER dataset that we use for evaluation.

E.5 Evaluation Tools and Methodologies

To evaluate our Named Entity Recognition (NER) tasks, we used a Perl script: conlleval.pl.

For other tasks, we calculated F1 score using Scikit-learn.

E.6 AI Assistants

We acknowledge the use of GPT-4 for grammar checking and assisting with coding throughout our research process.

Model	Top-Bottom	Top-Mid-Bottom	Left-Right	Left-Mid-Right	Wrapping	Inlay	Triple-Stack	Single
GPT-3.5 Few	23.1	22.00	20.14	15.56	9.74	14.29	7.14	21.00
GPT-3.5 Zero	24.01	16.00	25.17	2.00	3.59	0.00	0.00	57.00
GPT-4 Few	35.33	0.00	64.92	7.78	4.18	28.57	21.43	32.00
GPT-4 Zero	17.26	2.00	54.94	2.00	7.17	14.29	7.14	29.50
Ernie-Lite	21.70	12.00	52.20	2.00	7.17	14.29	66.67	67.50
Yi-6B	47.34	16.86	27.54	9.32	25.11	25.00	57.14	33.18
Qwen-7B	33.21	5.56	29.12	11.32	14.56	25.00	42.86	42.95
Baichuan-13B	35.27	11.38	22.45	3.44	28.34	25.00	42.86	37.12
Mistral-7B	27.48	14.56	33.45	12.34	30.43	25.00	28.57	51.46

Table 8: Accuracy of models across different structure types of Chinese characters.

Unicode	Character	Structure	Unicode	Character	Structure
U+4EBF	亿	LR	U+4ED9	仙	LR
U+4EC0	什	LR	U+4EE3	代	LR
U+4EC1	仁	LR	U+4EEA	仪	LR
U+4EC3	仃	LR	U+4EEB	仫	LR
U+4EC4	仄	WRP	U+4EF0	仰	LR
U+4EC7	仇	LR	U+4EF2	仲	LR
U+4ECE	从	LR	U+4EF5	仵	LR
U+4ED1	仑	TB	U+4EFB	任	LR
U+4ED3	仓	TB	U+4EFD	份	LR
U+4ED5	仕	LR	U+4F01	企	TB
U+4ED6	他	LR	U+4F0A	伊	LR
U+4ED7	仗	LR	U+4F0D	伍	LR
U+4ED8	付	LR	U+4F0E	伎	LR

Table 9: This table showcases a randomly selected range of Unicode characters in dataset along with their respective structures. This representation provides a snapshot of the structural information inherent in the Unicode.

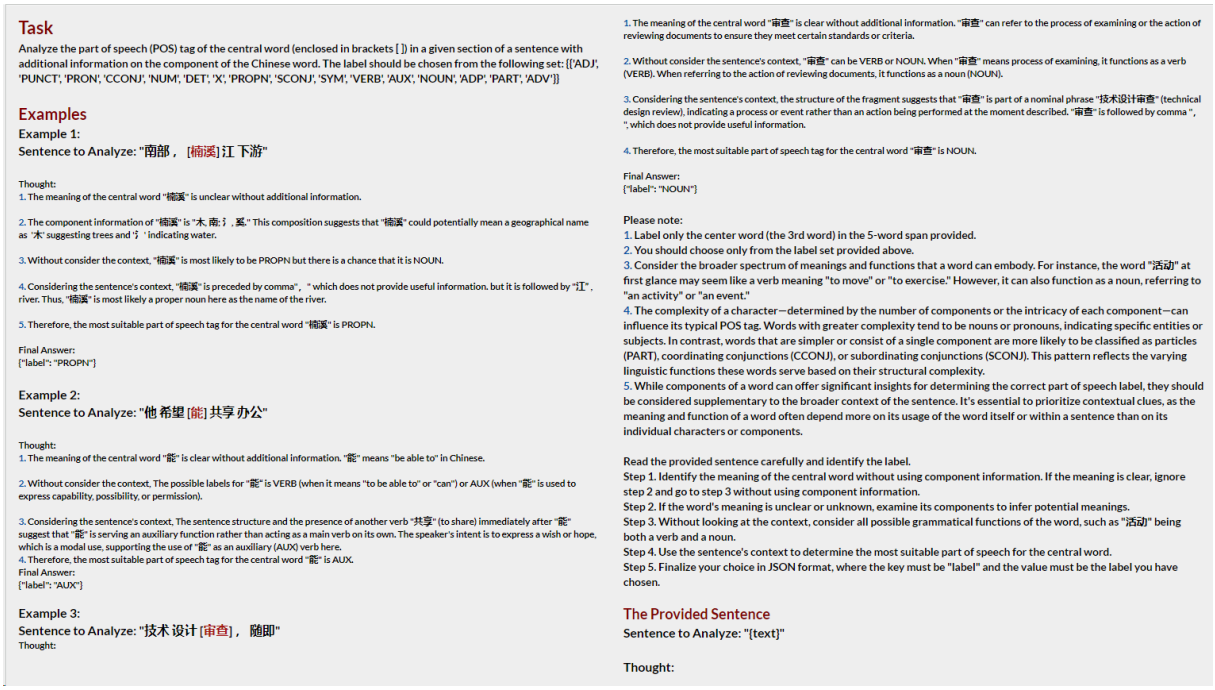

Figure 5: Radical Prompting for Chinese Part-of-Speech Tagging