

Morphology of Chinese Characters: Evaluating LLMs and VLMs on Visual Features and Radical Prompting for NLP Tasks

Anonymous ACL submission

Abstract

As a glyphic language, Chinese incorporates information-rich visual features below the character level, such as radicals which can provide hints about meaning or pronunciation. However, we argue that Large Language Models (LLMs) and Vision-Language Models (VLMs) fail to identify or harness these valuable features. Our study evaluates LLMs and VLMs in identifying visual information in Chinese characters, such as radicals, composition structures, strokes, and stroke count. Additionally, we design “radical prompting” to explore enhancements for LLMs in NLP tasks utilizing radical information. Results demonstrate most LLMs and VLMs struggle to recognize any visual information in Chinese characters. The introduction of ‘radical prompting’ led to some improvements in LLM performance across NLP tasks, but significant improvement was seen only when correct radicals were provided, as observed in part-of-speech (POS) tagging task.

1 Introduction

Unlike some of the alphabetical languages, a character is not the smallest meaningful unit in Chinese. Most Chinese characters are composed of simpler characters called radicals¹, which usually provide hints about the pronunciation or meaning of the character they form. Radicals and Chinese characters decompose into strokes, the smallest units. For example, the Chinese character “花” (meaning “flower”) has “艹” (meaning “herbal”) on the top, contributing to its semantic meaning, and “化” on the bottom, indicating its pronunciation. The component “化” can be further decomposed into “亻” and “匕.” Ultimately, everything can be decomposed into strokes as illustrated in Figure 1.

¹In its strictest sense, a radical refers to one specific building block of a Chinese character that can be used as a way of indexing; characters with the same radical share similar properties. In a broader definition, radicals refer to any large components within a character, which this paper adopts.

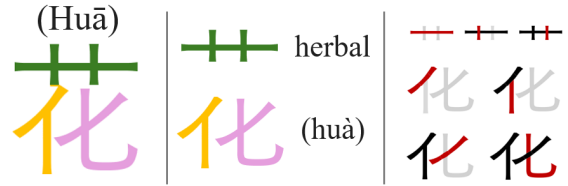


Figure 1: Chinese character “花” displayed at the character, radical, and stroke levels from left to right. Different radicals are shown in green, yellow, and pink colors, while the writing order of the strokes is indicated by red (current), gray (upcoming), and black (completed).

When encountering unknown characters, Chinese speakers rely on radicals to estimate their meanings or pronunciations and strokes to know how to write them. Therefore, we evaluated LLMs and VLMs on their ability to identify these visual features and designed ‘radical prompting’ to adopt similar strategies in NLP tasks.

To evaluate models’ ability to recognize visual information embedded in Chinese characters, we constructed a dataset including over 14,000 Chinese characters from CJK Unified Ideographs², annotated with three features: radicals, strokes, and stroke count. Among these, 4,564 Simplified Chinese characters are also annotated with composition structure information. Radicals provide clues to the meaning or pronunciation of characters. Composition structures influence the order and method of recognizing and decomposing characters. Strokes dictate the sequence of writing a character, and stroke count offers a measure of a character’s visual complexity and density.

Results demonstrate LLMs and VLMs—whether Chinese or multilingual—struggle to accurately identify visual features in Chinese characters. Mod-

²Modern typefaces for Japanese and Korean languages often utilize their regional variations of Chinese characters. The CJK Unified Ideographs refers to a set of Chinese characters used across these three languages, where efforts have been made to standardize and unify the characters to accommodate their shared use in digital typography and computing.

els display low accuracy and F1 scores, and high entropy when asked to identify radicals, structure, or strokes of Chinese characters. They also show high mean squared error (MSE) in estimating stroke counts. When utilizing “radical prompting” in NLP tasks, which asks models to utilize radical information on unfamiliar words or characters, LLMs show improvement when correct radicals are given. However, when not given the correct radicals, the improvements are minimal.

This paper makes three key contributions to improve Chinese language processing in LLMs and VLMs at the character level and below: 1) It develops a dataset that captures the visual aspects of Chinese characters; 2) examines the challenges contemporary LLMs face in precisely recognizing the visual information of characters; 3) introduces ‘radical prompting’ to enhance the performance of LLMs on general NLP tasks.

2 Related Work

Chinese Character Decomposition in Computer Vision

The task of decomposing Chinese characters into constituent components aligns with the field of computer vision. Research within this domain, such as the studies by (Ma et al., 2021), (Xia, 1994), and (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 which we adopt in this paper.

Chinese decomposition dataset In reviewing available resources, we encountered a comprehensive dataset (Kawabata et al.) that offers decompositions for the CJK Unified Ideographs. Although this collection overlaps with our dataset, it does not cite any authoritative sources for its data. This omission leads to ambiguity due to multiple decomposition sequences for individual characters.

Our approach utilizes sources from authoritative dictionaries such as the Kangxi Dictionary (康熙字典) and the Xinhua Dictionary (新华字典)³, ensuring a validated framework for visual information. Additionally, our dataset contains systematic and standard stroke orders for all 14,648 characters,

³Xinhua and Kangxi Dictionaries are renowned lexicographical resources for Chinese. Digitalized Kangxi Dictionary can be found here: <https://www.kangxizidian.com/>

which the aforementioned dataset lacks. We also created a manageable subset of 4,651 Simplified Chinese characters with structural classification.

Glyphic Embedding Strategies in LMs Recent studies have sought to leverage the rich visual information inherent in Chinese characters to enhance language model performance. For instance, (Sun et al., 2021) introduces a novel approach that incorporates different embeddings alongside glyph embeddings derived from various fonts to enrich character representations. Similarly, (Si et al., 2021) delves into the potential of stroke encoding among other glyph-based input methods to explore their performance. Additionally, (Shi et al., 2015) harnesses radical information, utilizing it as a key component for embedding Chinese characters.

These approaches 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. Tokenizing at a fine-grained level also leads to slower output speeds. Our paper, in contrast, zeroes in on the impact of incorporating visual features of Chinese characters directly within contemporary LLMs and VLMs.

Statistic	Number
Total Characters	14,648
- Frequently used :	3,500 (24.1%)
- Commonly used :	3,000 (20.6%)
- Terminology used:	1,605 (11.0%)
- Rarely used:	5,543 (37.8%)
- With structural information:	4,651 (31.8%)
Without components	324
With 2 components	12,769
With 3 components	992
With more than 3 components	476
Unique stroke patterns	13,740
Stroke number mean	11.51
Stroke number standard deviation	3.92
Minimum stroke number	2
Maximum stroke number	39

Table 1: Key statistics of our Chinese character dataset

3 Dataset

To evaluate contemporary LLMs and VLMs’ proficiency with visual information in Chinese characters, we compiled a dataset using characters from CJK Unified Ideographs with visual features collected from the digitized Kangxi Dictionary (康熙

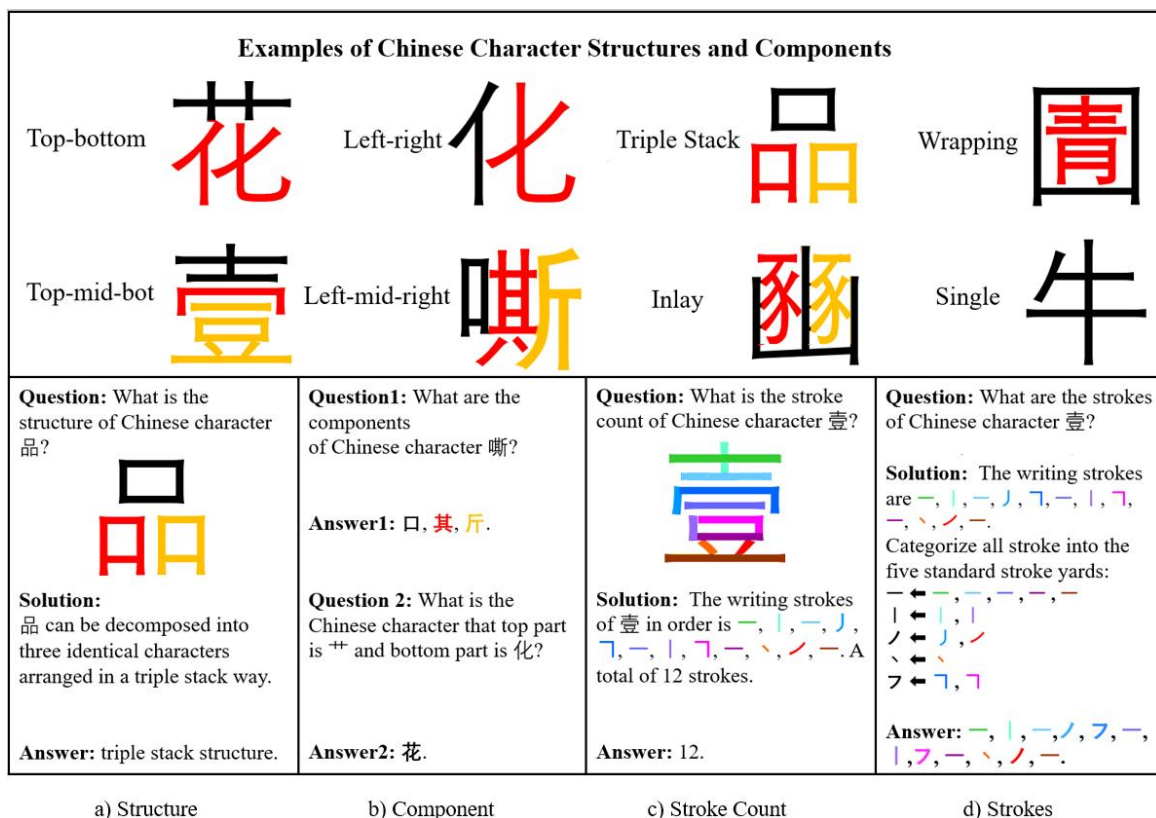


Figure 2: Examples of composition structures with radical in order of black, red, yellow and four types of tasks.

字典) and Xinhua Dictionary (新华字典). Our dataset includes 14,648 Chinese characters and details their corresponding radicals, strokes, and stroke count. A subset of 4,651 Simplified Chinese characters also contains structural composition information. The detailed statistics are provided in Table 1 with three tiers of Chinese character frequency listed for reference. These tiers are categorized by the Table of General Standard Chinese Characters published by the Chinese government.

Structure of Chinese Characters. According to the digitized Kangxi dictionary, we categorize 4651 simplified Chinese characters into eight major structural arrangements: top-bottom, left-right, top-mid-bottom, left-mid-right, wrapping, inlay, triple-stack, and single structure, which refers to characters that cannot be further segmented. Examples of each structure are illustrated in Figure 2, with the distribution of structure types shown in Figure 3. The structure of Chinese characters can be complex, with layers of structure compounding upon each other. For example, the character ‘花,’ shown in Figure 1, has a top-bottom structure, consisting of “+” and “化.” “化” exhibits a left-right

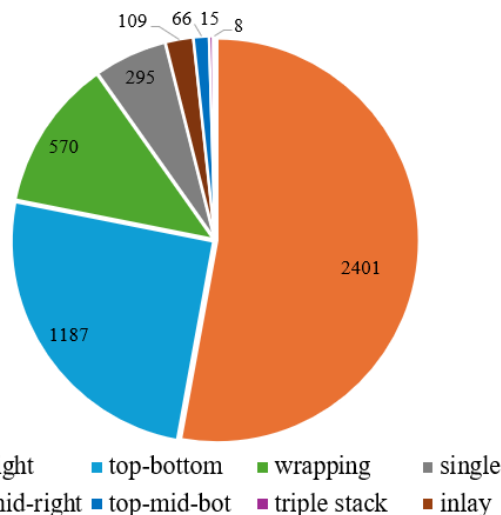


Figure 3: The proportion of Chinese character structures

structure which can be further decomposed into “亻” and “七.” To maintain clarity, we categorize all characters based on their primary structure.

Radicals of Chinese Characters. Radicals are the major component blocks in Chinese characters, providing essential clues about meaning and pronunciation. In our dataset, the radicals were

collected using a combination of human annotation and APISpace’s Chinese character segmentation API⁴. After attempts at automated annotation, we manually reviewed and adjusted segmentations to ensure that at least one component is meaningful after segmentation, wherever feasible. For example, while “八” could be segmented as a left-right structure, we classify “八” as a single structure with zero radicals to avoid all radicals being meaningless strokes after segmentation. Approximately 1,000 characters required manual adjustment due to empty or incorrect radicals, with more than 500 being adjusted to avoid reduction to strokes by one of the authors who is a native Chinese speaker.

The radical order follows rules: from top to bottom, left to right, outside to inside, and main part before inlay parts as illustrated in Figure 2, where the radicals are colored according to their order and structures. If a radical does not exist in the typeface, it is further split to check for existing sub-radicals. For example, in a left-mid-right structured character, if the mid part cannot be typed but can be split into top and bottom parts, the radical order will be left, mid (top), mid (bottom), and right.

Strokes of Chinese Characters. Chinese dictionaries categorize all Chinese strokes into five basic stroke types: “一”, “丨”, “丿”, “丶”, and “フ”, which our dataset adopts. We first utilized the Xinhua Dictionary (新华字典) API to annotate the strokes. For characters not found in the dictionary, we attempted to concatenate the stroke information of their components in order. We then manually reviewed the stroke information to ensure accuracy.

The stroke count, also collected in the dataset, is the number of strokes required to write a character, offering a measure of word complexity and density. Unlike alphabetic languages, where word length can hint at complexity, Chinese characters occupy uniform space, making stroke count a valuable indicator of intricacy. The statistics for strokes are provided in Table 1 with illustrations in Figure 2.

4 Evaluation on Visual information of Chinese Character

4.1 Tasks

Structure Recognition of Chinese Characters. We assess LLMs and VLMs’ proficiency in identifying the correct structural arrangements of Chi-

nese characters. For this task, we provide the character along with eight different structure types and ask the model to identify which type it is. Performance is evaluated using the F1 score.

Radical Recognition of Chinese Characters. We evaluate LLMs and VLMs’ capability to recognize radical information of Chinese characters in two ways: character-to-radical and radical-to-character. In the first type, models are expected to output the character’s radical in the correct order, which implicitly requires structural knowledge. Performance is assessed by the accuracy of the first three radicals and the overall F1 score, measuring the recognized order and the number of radicals recognized, respectively. In the second type, models are provided with radical information along with their relative positions based on structures, and they are expected to identify the correct characters. Performance is evaluated by accuracy.

Stroke Count Identification of Chinese Characters. We measure the LLMs and VLMs’ effectiveness in determining the stroke count of Chinese characters. Models are tasked with identifying the total number of strokes required to write each character. Performance is measured using Mean Absolute Error (MAE) and Mean Squared Error (MSE).

Stroke Identification of Chinese Characters. Similar to radical recognition, we evaluate LLMs and VLMs’ ability to identify the sequence of strokes required to write a character. Performance is calculated using the overall F1 score, with positional accuracy for the first three positions.

4.2 Experimental Setup

We evaluate the visual information of Chinese characters using various LLMs and VLMs. The multilingual LLMs include Aya (Üstün et al., 2024), Claude-3 (Anthropic, 2024), Gemini-1.5, GPT-3.5 Turbo (OpenAI, 2024) in zero-shot, few-shot, Chain-of-Thought (CoT), and fine-tuning settings, and GPT-4 (OpenAI, 2023). The Chinese-English bilingual LLMs include ERNIE-Lite (Baidu, 2024a), Kimi-v1 (MoonshotAI, 2024), and several open-source LLMs such as Baichuan-13B (BaichuanInc, 2024), BLOOM-7B (BigScience, 2024), ChatGLM-6B (Zeng et al., 2023), ChineseLLaMA-7B (HFL, 2024), InternLM-7B (InternLM, 2024), Orion-14B (Chen et al., 2024), Qwen-7B (Bai et al., 2023), Qwen-2-72B, and Yi-6B (AI et al., 2024). We also evaluate VLMs providing im-

⁴API document in Chinese can be accessed through link: <https://www.apispace.com/eolink/api/dfsdfsfsf/apiDocument>

Model	Structure		Radicals						Stroke Count		Strokes						
	F1	H	1st		2nd		3rd		F1	Acc	MSE	MAE	1st	2nd	3rd	H	F1
			Acc	H	Acc	H	Acc	H					Acc	Acc	Acc		
<i>Close-Sourced Models (LLMs)</i>																	
Aya	12.56	0.16	35.72	0.45	2.16	0.86	0.26	0.89	20.13	5.65	13.20	2.79	28.24	23.48	19.44	0.37	21.43
Claude-3	23.70	0.54	70.02	0.37	5.64	1.43	0.43	1.46	45.57	40.40	7.78	1.32	28.64	19.02	31.19	0.88	22.91
Gemini-1.5	23.04	0.56	4.20	1.00	0.04	1.72	0.38	0.77	1.37	11.26	13.23	2.76	26.66	24.52	15.14	0.81	20.24
Few-shot GPT-3.5	22.82	0.88	54.14	0.49	7.37	1.44	0.30	1.70	34.60	23.12	7.96	1.65	27.86	22.70	30.23	1.13	25.62
Zero-shot GPT-3.5	15.43	0.69	52.14	0.71	4.33	1.65	0.20	1.54	31.66	17.45	48863	5.99	30.70	21.92	26.97	0.98	25.09
Fine-tune GPT-3.5	27.14	0.33	4.12	0.50	0.00	1.65	0.00	1.18	1.23	71.66	7.36	1.46	47.50	44.58	32.67	1.08	28.64
CoT GPT-3.5	38.08	1.25	5.24	0.57	0.16	1.68	0.11	0.91	1.63	24.41	8.93	1.92	31.06	22.22	26.85	0.83	25.60
Few-shot GPT-4	45.28	0.48	58.44	0.58	6.45	1.29	0.31	0.65	41.66	38.01	7.96	1.65	24.18	18.22	21.90	1.37	20.87
Zero-shot GPT-4	35.40	0.54	57.86	0.63	6.28	1.37	0.20	0.63	41.42	38.76	12.17	1.99	27.04	21.16	21.99	1.21	22.18
Ernie-Lite-8K	7.19	0.76	18.92	1.68	3.52	2.00	0.13	1.98	11.99	3.72	44.53	5.34	29.30	23.28	20.78	1.11	23.34
Kimi-v1	24.51	0.83	7.24	0.48	0.33	1.05	0.00	0.62	1.10	50.16	19.05	3.12	33.12	21.56	19.72	1.07	22.99
<i>Open-Sourced Models (LLMs)</i>																	
Baichuan-13B	11.17	0.88	33.20	0.98	2.05	1.65	0.60	0.98	22.62	13.67	32.70	4.31	27.68	21.42	15.92	1.56	22.74
Bloom-7B	9.81	0.96	3.48	1.71	0.54	2.12	0.04	1.28	4.15	0.00	46.76	4.05	27.92	24.96	14.47	0.87	23.19
ChatGLM-6B	10.30	0.68	6.94	1.56	0.50	1.67	0.00	0.82	6.33	1.38	29.68	4.25	26.88	12.60	12.43	0.96	27.28
Chinese-LLaMA-7B	5.13	0.97	9.26	1.89	0.64	2.10	0.17	1.78	6.32	0.32	15.83	3.00	26.26	24.86	13.42	0.93	22.32
InternLM-7B	9.68	1.05	12.08	1.71	0.34	1.68	0.05	1.11	8.89	0.00	45.38	5.50	28.82	24.66	13.38	0.95	22.01
Orion-14B	9.00	1.04	5.27	1.73	0.18	1.14	0.76	0.47	9.46	3.39	31.45	4.45	28.40	22.82	19.38	0.90	24.81
Qwen-7B	5.25	1.16	17.30	1.54	0.85	1.82	0.23	1.14	12.41	1.59	34.16	4.62	25.02	20.20	21.92	1.30	23.30
Qwen-2-7B	6.76	1.50	15.42	1.74	0.68	1.99	0.22	1.53	10.70	0.42	44.48	5.39	23.16	18.50	21.54	1.40	22.68
Yi-6B	8.86	0.70	14.18	1.60	1.05	1.80	0.21	0.81	12.14	0.32	29.49	4.24	28.56	22.40	7.76	0.85	24.17
<i>Vision Language Models (VLMs)</i>																	
Claude-3V	23.70	0.54	8.80	0.39	0.61	1.49	0.00	1.38	2.44	57.30	5.93	1.22	15.40	19.60	26.80	1.22	19.62
Ernie-4V	54.30	-	41.03	-	34.21	-	12.50	-	41.67	71.79	12.54	1.78	53.85	35.90	47.37	-	30.90
Gemini-1.5V	27.15	0.36	3.00	0.76	0.41	1.41	0.00	1.20	1.53	27.08	8.83	2.28	29.60	16.80	22.00	1.00	22.04
GPT-4V	23.28	0.46	10.20	0.57	0.41	1.50	0.00	0.77	9.22	24.18	7.96	1.64	24.00	19.60	23.80	1.34	21.96
Kimi-V	45.60	-	36.73	-	19.15	-	0.00	-	32.93	42.86	15.32	2.68	30.61	26.53	16.67	-	20.70
Fine-tune PIXEL	84.57	-															

Table 2: Models Performance on Chinese character visuals with each task separated by vertical lines. **H**: Entropy. The top scores for each section and overall are highlighted in blue and red respectively.

ages of characters in Microsoft YaHei⁵ font, including Claude-3V, Ernie-4V (Baidu, 2024b), Gemini-1.5V, GPT-4V, and Kimi-V. Additionally, we assess the pixel-based encoder model, PIXEL (Rust et al., 2023). Since PIXEL is a language model lacking sentence completion abilities, it is only evaluated on the structure recognition task using a span-based question-answering framework after fine-tuning.

For our evaluation, we use different sampling methods and settings based on the type of model. For LLMs, a random sample of 1,000 characters is selected for each task and model. Due to higher costs, the number of samples for VLMs is reduced to 500. ERNIE-V and Kimi-V, which lack API access, are tested manually with only 100 samples. We incorporate few-shot learning by providing models with three examples for each task, except for the structure recognition task, where one example per structure type is given. In the Chain-of-Thought (CoT) setting, models are prompted to break down their reasoning process step-by-step,

⁵Yahei is the default Chinese font in Microsoft Office.

with detailed prompts provided in the Appendix B.2. Models with fine-tuning are trained with a 7:3 split and tested using 1,000 samples randomly selected from the test set. To assess consistency and model entropy, each question is asked five times, and the best trial out of the five for each task is selected to calculate the overall results.

To adapt answers from models generating long responses conventionally, we first let models generate responses freely without a specific answer format. Then, we use GPT-3.5 Turbo to extract answers from various model responses. For open-source models and extraction-used GPT-3.5 Turbo, a temperature of 0.3 is applied. Closed-source models generally use a temperature of 0.7 unless otherwise recommended by model documentation.

4.3 Experimental Result

As illustrated in Table 2, the performance of the evaluated models on Chinese character visual information tasks is generally low across the board. For the structure recognition task, only PIXEL

achieved a high score of 84.57. In contrast, the majority of other models scored below 30. In the radical-to-character task, a clear trend emerges where model performance is highest for the first component and sharply decreases for subsequent components. For example, GPT-3.5 in zero-shot achieved an F1 score of 52.14 for the first component, but this dropped to 4.33 for the second component and nearly zero for the third. This phenomenon is also reflected in the increase of entropy along positions, suggesting that models possess some vague knowledge of characters, particularly the attribute radical that often appears first, such as “艹” in “花”. For the stroke count identification task, the majority of models achieved a MSE higher than the average stroke number in the dataset: 11.51. Accuracy scores for stroke identification remained consistent across positions. However, the overall performance remained poor with Ernie-4V achieving the highest overall f1 of 30.90.

Although fine-tuning, Chain-of-Thought prompting, and multilingual VLMs led to some improvements on various tasks, we observed a significant drop in performance for the character-to-component task. For instance, the overall F1 score dropped from 45.57 to 2.44 when Claude-3 is provided with images. Conversely, Chinese models such as Ernie-V showed substantial improvement with images. In the component-to-character task, Ernie-V achieved the highest accuracy at 71.79% among all, relying heavily on image information to extract characters for the answer and failing whenever it misrecognized the character in the image.

There are several types of characters that Ernie-V and Kimi-V tend to make mistakes on. Firstly, complex and dense characters are often misrecognized as similar, more frequently used characters. In a uniform space, as characters become more complex, the individual radicals within the character become narrower, leading to misrecognition. Secondly, characters that are extremely similar, with only a single stroke difference, are often seen by the models as the more common variant of the two. Thirdly, for rare characters, Ernie-V often states that it does not detect any character in the image, while Kimi-V even refuses to allow the user to send the prompt when it fails to extract the character from the image. Models occasionally recognize a radical of the character as the character itself. They sometimes confuse the character in the image with black and white pictures. Examples of Kimi-V and Ernie-V’s behavior are provided in Appendix B.3.

Q: 南部, [楠溪] 江下游
What is the POS tagging of [楠溪]?
A:
○ 楠溪 is unfamiliar.
○ 楠 is composed of 木, 南.
• 木 means wood.
• 南 means south.
溪 is composed of 氵, 奚.
• 氵 refers water.
○ 南溪 could be a geographical name...
○ The tag should be PROP.

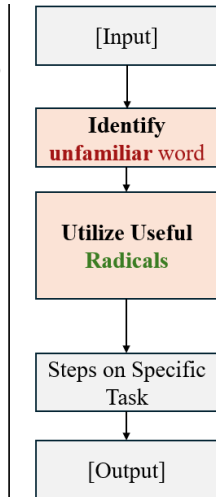


Figure 4: Example of model answer for part-of-speech (POS) tagging with an unfamiliar Chinese word using radical prompting. **Gray blocks** represent steps that also appear in the baseline procedure, while **light orange blocks** indicate the unique steps in radical prompting.

Although improvement is shown in VLMs and other methods, the overall performance remains poor compared to a native Chinese speaker, who would achieve nearly full scores for all tasks even without knowing the majority of the queried characters. To further explore visual information in text input, we experimented with the impact of Chinese character encodings on visual tasks. Details and results are in Appendix C.

5 Evaluation on Radical Prompting

In this section, we examine the significance of the most important visual information in Chinese characters: radicals. Similar to the strategy used by Chinese speakers when encountering unfamiliar words, we prompt models to use their knowledge of radicals to estimate the meaning of the characters when they encounter unfamiliar terms. Examples of this prompting strategy are shown in Figure 4.

5.1 Tasks

Part-of-Speech (POS) tagging. We utilized three datasets for this task: the GSD Simplified dataset (Qi and Yasuoka, 2023), the Parallel Universal Dependencies (PUD) dataset (McDonald et al., 2023), and a new dataset we annotated using Classical Chinese RoBERTa (Yasuoka, 2023) with 500 sentences from Tang Dynasty poems⁶. For this

⁶The choice of classical poems is motivated by the compactness of information in poetry, suggesting more information is preserved at the sub-character level.

Model	Part-Of-Speech Tagging								
	GSD			PUD			Poems		
	B	RP	RP (Oracle)	B	RP	RP (Oracle)	B	RP	RP (Oracle)
Aya	68.86	68.91(+0.1)	70.41(+1.6)	73.87	77.21(+3.34)	76.95(+3.08)	65.53	66.19(+0.66)	66.71(+1.18)
Claude-3	69.37	70.68(+1.3)	70.45(+1.1)	69.37	70.45(+1.08)	70.68(+1.31)	65.53	66.20(+0.67)	66.71(+1.18)
ERNIE-Lite-8K	27.06	24.97(-2.1)	32.73(+5.7)	30.35	30.29(-0.06)	41.29(+10.94)	44.19	42.17(-2.02)	49.07(+4.88)
GPT-3.5	59.08	64.62(+5.5)	67.56(+8.5)	62.61	69.90(+7.29)	73.46(+10.85)	53.51	59.22(+5.71)	61.39(+7.88)
GPT-4	71.55	72.14(+0.6)	72.95(+1.4)	76.20	76.72(+0.52)	77.35(+1.15)	66.94	67.11(+0.17)	67.57(+0.63)
QWen-72B	62.20	65.38(+3.2)	67.32(+5.1)	62.20	65.38(+3.18)	67.32(+5.12)	55.63	57.78(+2.15)	59.54(+3.91)

Table 3: Model performance for POS tagging with baseline(B), radical prompting without golden components (RP), and radical prompting with oracle information (RP (Oracle)). Performance change relative to baseline is highlighted with green for increase and red for decrease.

Models	Name Entity Recognition				Chinese Word Segmentation			
	People’s Daily		Weibo		GSD		PUD	
	B	RP	B	RP	B	RP	B	RP
Aya	38.24	36.36(-1.9)	37.88	30.83(-7.05)	87.98	89.08(+1.1)	88.68	91.05(+2.37)
Claude-3	69.74	73.79(+4.1)	45.64	46.86(+1.22)	94.90	95.16(+0.3)	94.12	94.96(+0.84)
ERNIE-Lite	12.10	12.99(+0.9)	6.72	6.90(+0.19)	88.04	88.70(+0.3)	69.54	73.57(+4.03)
GPT-3.5	56.89	55.97(-0.9)	36.65	36.64(-0.01)	95.68	94.87(-0.8)	93.91	93.70(-0.21)
GPT-4	66.04	68.05(+2.0)	43.83	44.68(+0.85)	94.21	94.88(+0.7)	94.24	95.63(+1.39)
QWen 72B	62.73	59.59(-3.1)	31.78	35.83(+4.05)	96.59	95.57(-1.0)	89.79	91.94(+2.15)

Table 4: Model performances for NER and CWS tasks with baseline(B) and radical prompting(RP).

task, we used a 5-word span to focus on predicting single word with or without radical information. Results on different word span sizes are shown in Appendix D.2. Word spans containing more than one punctuation mark were excluded. We designed two versions of the task: the first supplies the correct component and radical information of the central word, while the second version prompts the model to utilize radical without explicitly providing it. The models’ performance on this task is measured using the F1 score.

Named Entity Recognition (NER). We use two distinct datasets for the NER task: 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. We dropped tagging on nominal entities and focused on three types of tags: PER (person), LOC (location), and ORG (organization), with BIO tagging standard. Due to the comprehensive understanding required for the NER task, providing radical information for each word and character would overwhelm the model. Therefore, we evaluate the efficacy of radical prompting by guiding the model to use radicals, rather than supplying all the radical information. Answers are evaluated using the F1 score.

Chinese Word Segmentation (CWS). For this task⁷, we give whole sentences from the GSD and PUD datasets and ask models to separate them into words. Answers are evaluated using the F1 score.

5.2 Method

Baseline. Our baseline employs the Chain-of-Thought (CoT) prompting framework with steps that guide the model to execute tasks.

Radical Prompting. We introduce a novel prompting method named “radical prompting,” which builds upon the CoT framework. The process begins with the model identifying any unclear words within a given context. Then, the model is instructed to dissect these words into their constituent radicals and attempt to utilize useful radicals to aid the task. Steps are then provided to guide the model in executing specific tasks, identical to the baseline, with three examples. When using radical prompting, it is important to guide models to critically assess information from character components to avoid being misguided. Thus, one example intentionally includes radical infor-

⁷CWS 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 could be beneficial.

mation that is irrelevant. Prompt lines of radical prompting are listed in Appendix D.3.

5.3 Experimental Setup

We apply a similar strategy for model selection and sample number as Section 4.2. The models selected for evaluation are Aya, Claude-3, ERNIE-Lite-8K, GPT-3.5, GPT-4, and QWen-1.5 72B Chat. Each task and each dataset uses 2,000 sample sentences, repeated 5 times for consistency. Due to higher costs, Claude-3 and GPT-4 are evaluated using only 1,000 sample sentences. Models are asked to return answers in JSON format with target sentences annotated similar to (Blevins et al., 2023).

5.4 Experimental Result

We observed varying performance impacts on the two tasks: when context is limited, such as in POS tagging, information embedded in characters becomes beneficial. For the POS tagging task, performance generally improved across models, except ERNIE-Lite when correct radicals were not provided. Most models showed significant improvement when correct radicals were supplied. For example, on the PUD dataset, GPT-3.5 improved by 7.29 points without correct radicals but gained an additional 10.85 points when correct radicals were provided. A qualitative analysis of the impact of radicals on POS tagging is provided in the Appendix D.1. For the NER and CWS tasks, we observed a mixture of positive and negative impacts after applying radical prompting. While Claude-3 and GPT-4 show consistent improvement across datasets and tasks, models such as Aya exhibit a noticeable drop of 7.05 points in the Weibo NER dataset after applying radical prompting. This variation may be attributed to the inherent difficulty of the tasks. Upon reviewing the models’ responses, we found that radicals tend to misguide the models when dealing with relatively new terms, such as transliterated foreign terms—terms that have been adapted to Chinese words based solely on pronunciation. This misguidance likely contributed to the mixed results observed.

6 Discussion on Chinese Characters

To investigate the importance of Chinese radicals, we selected a sample of 100 Chinese characters from our dataset and annotated them to determine whether the radicals directly contribute to the meaning or pronunciation of the character, as shown in

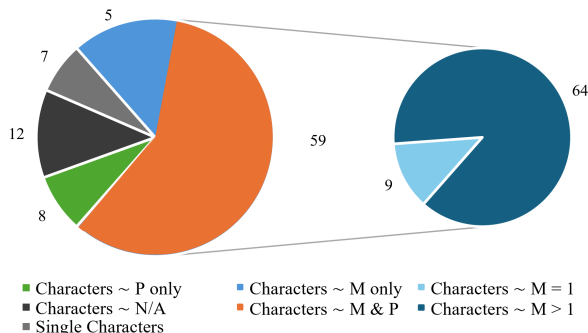


Figure 5: Distribution of Chinese characters with meaning (M) or pronunciation (P) hint from their radicals. The smaller circle on the right shows the distribution among all characters containing radicals with meaning (sum of Characters M only and Characters M & P).

Figures 5. Although the majority of characters have clues derived from the radicals, we found that most characters contain a combination of only one meaningful radical with other radicals hinting at pronunciation. For example, in the character “花,” we can infer that it is related to herbs from the radical “艹,” while “化” only provides a pronunciation hint, resulting in only vague idea of character’s meaning. In 12 out of the 100 characters, none of the radicals were helpful. This is due to the evolution of the language, where historically, a single Chinese character often conveyed the meaning of a full word. However, more words are now composed of two or more characters, leading to individual characters losing their original meanings. For example, the Chinese character “况” is now commonly used to mean “situation” in words like “情况” or “状况”. However, the original meaning of the character is “cold water” unexpectedly, which is closely related to the radical “冫”, referring to cold water.

7 Conclusion

In this paper, we create a comprehensive dataset on visual information in Chinese characters and evaluate the ability of LLMs and VLMs to recognize this information. Our observations highlight the suboptimal performance of these models in handling information below the character level. Despite this, our experiments with ‘radical prompting,’ which prompts models to utilize radical information, demonstrate that these sub-character features can still be beneficial. The results show stable improvements in POS tagging, and mixed results of improvement and decline for NER and CWS without correct radicals provided.

520 Limitations

521 Our study, while contributing valuable insights into
522 the integration of radical prompting for Chinese lan-
523 guage models, encounters several limitations that
524 suggest directions for future research. First, the
525 dataset employed does not encompass the full array
526 of Chinese characters but is confined to commonly
527 used characters. This selective coverage might af-
528 fect the scalability of our findings to all Chinese
529 characters especially when greater model meets un-
530 known or unfamiliar character, there is a chance
531 that our dataset does not cover that character.

532 Additionally, the study primarily evaluates the
533 effectiveness of radical prompting on a narrow se-
534 lection of models and specific NLP tasks, which
535 might not reflect its utility across different models
536 or broader language processing applications.

537 Furthermore, an intrinsic limitation of our
538 methodology arises from the exclusive use of En-
539 glish in our prompting lines. Incorporating Chinese
540 in the prompting strategy could potentially enhance
541 the relevance and effectiveness of prompts, align-
542 ing better with the linguistic context of the target
543 language.

544 References

545 01. AI, :, Alex Young, Bei Chen, Chao Li, Chen-
546 gen Huang, Ge Zhang, Guanwei Zhang, Heng Li,
547 Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong
548 Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang,
549 Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang,
550 Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng
551 Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai,
552 Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024.
553 [Yi: Open foundation models by 01.ai.](#)

554 Anthropic. 2024. [Introducing the claude 3 family.](#) Ac-
555 cessed: 2024-06-09.

556 Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang,
557 Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei
558 Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin,
559 Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu,
560 Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren,
561 Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong
562 Tu, Peng Wang, Shijie Wang, Wei Wang, Sheng-
563 guang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang,
564 Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu,
565 Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingx-
566 uan Zhang, Yichang Zhang, Zhenru Zhang, Chang
567 Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang
568 Zhu. 2023. Qwen technical report. *arXiv preprint*
569 *arXiv:2309.16609*.

570 BaichuanInc. 2024. [Baichuan-13b-base.](#) Accessed:
571 2024-06-11.

Baidu. 2024a. [Introducing ernie 3.5: Baidu’s
knowledge-enhanced foundation model takes a gi-
ant leap forward.](#) Accessed: 2024-06-11. 572
573
574

Baidu. 2024b. [Yiyan.](#) Accessed: 2024-06-11. 575

BigScience. 2024. [Bloom-7b1.](#) Accessed: 2024-06-11. 576

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. [Prompting language models for linguistic structure.](#)
In *Proceedings of the 61st Annual Meeting of the
Association for Computational Linguistics (Volume
1: Long Papers)*, pages 6649–6663, Toronto, Canada.
Association for Computational Linguistics. 577
578
579
580
581
582

Du Chen, Yi Huang, Xiaopu Li, Yongqiang Li,
Yongqiang Liu, Haihui Pan, Leichao Xu, Dacheng
Zhang, Zhipeng Zhang, and Kun Han. 2024. [Orion-
14b: Open-source multilingual large language mod-
els.](#) *arXiv preprint arXiv:2401.12246*. 583
584
585
586
587

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 tag-
ging strategy. License: CC0: Public Domain. 588
589
590
591
592
593
594

HFL. 2024. [Chinese llama-2-7b.](#) Accessed: 2024-06-
11. 595
596

InternLM. 2024. [Internlm-7b.](#) Accessed: 2024-06-11. 597

Kawabata Kawabata, Masaya Nakamura, and Huáng
Jùnliàng. Cjkvi-ids: Ideographic description se-
quences for cjk unified ideographs. [https://
github.com/cjkvi/cjkvi-ids](https://github.com/cjkvi/cjkvi-ids). Accessed: 2024-
4-4. 598
599
600
601
602

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. 603
604
605
606
607

Jiefeng Ma, Zirui Wang, and Jun Du. 2021. [An open-
source library of 2d-gmm-hmm based on kald toolkit
and its application to handwritten chinese character
recognition.](#) *Lecture Notes in Computer Science*,
12888. 608
609
610
611
612

Ryan McDonald, Joakim Nivre, Yvonne Quirnbach-
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 dependen-
cies \(pud\) treebanks for multilingual parsing.](#) Avail-
able 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. 613
614
615
616
617
618
619
620
621
622

MoonshotAI. 2024. [Kimi.](#) Accessed: 2024-06-11. 623

OpenAI. 2023. [Gpt-4.](#) Accessed: 2024-06-11. 624

625	OpenAI. 2024. Gpt-3.5 turbo . Accessed: 2024-06-11.	Jianshu Zhang, Yixing Zhu, Jun Du, and Lirong Dai. 2018. Radical analysis network for zero-shot learning in printed chinese character recognition. In <i>Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)</i> , Hefei, Anhui, P.R. China. IEEE.	679 680 681 682 683 684
626	Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In <i>Proceedings of the Human Language Technology Center of Excellence</i> , Baltimore, MD. Johns Hopkins University.	Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. <i>arXiv preprint arXiv:2402.07827</i> .	685 686 687 688 689 690 691 692
631	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.	A General Experiment Details	693
635	Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels . In <i>The Eleventh International Conference on Learning Representations</i> .	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:	694 695 696 697 698
640	Xinlei Shi, Junjie Zhai, Xudong Yang, Zehua Xie, and Chao Liu. 2015. Radical embedding: Delving deeper into chinese radicals. In <i>Proceedings of the Association for Computational Linguistics (ACL)</i> . Sogou Technology Inc., Beijing, China.	<ul style="list-style-type: none"> • GPT-3.5 and GPT-4 were used with the snapshot dated <i>2023-11-06</i>. • Claude model’s evaluation utilized the <i>2024-02-29</i> snapshot. • Ernie-Lite-8K was tested using the <i>2023-09-22</i> snapshot. 	699 700 701 702
645	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 . <i>Transactions of the Association for Computational Linguistics</i> , 9:634–649.	Temperature Settings	705
651	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 <i>Proceedings of the Association for Computational Linguistics (ACL)</i> . Shan-non.AI; Zhejiang University; Key Lab of Intelligent Information Processing of Chinese Academy of Sciences.	<ul style="list-style-type: none"> • Aya, Yi-6B, Qwen-7B-Chat, Baichuan-13B, and Mistral-7B were set at a lower temperature of <i>0.3</i> as recommended. • For other models not specifically mentioned, a temperature setting of <i>0.7</i> was used. 	706 707 708 709 710
659	Franck Xia. 1994. Knowledge-based sub-pattern segmentation: decompositions of chinese characters . <i>Proceedings of the International Conference on Image Processing</i> .	B Details on Visual Info Evaluation	711
663	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.	B.1 Structure Recognition Across Structures	712
671	Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model . In <i>The Eleventh International Conference on Learning Representations (ICLR)</i> .	We provide detailed result for structure recognition across different structures in Table 5.	713 714
672		B.2 CoT Prompting	715
673		We present the prompt lines used in Figure 6, 7, 8, 9.	716 717
674		B.3 Chinese VLMs Behavior	718
675		Examples of VLMs misrecognizing images are shown in Fig 10, 11, 12, 13, and 14.	719 720

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 5: 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	佗	LR
U+4EC4	仄	WRP	U+4EF0	仰	LR
U+4EC7	仇	LR	U+4EF2	仲	LR
U+4ECE	从	LR	U+4EF5	件	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 6: 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.

The structure of a Chinese character must be one of the following:
 上下结构, 左右结构, 上中下结构, 左中右结构, 包围结构, 镶嵌结构, 单一结构, 品字结构.
Let's think step by step. First identify the radical of the character. The radical is usually associate 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.
 The structure of Chinese character {character} is

Figure 6: Prompt Line of Structure Task

C Analysis on Chinese Encoding

The difficulty of the task is primarily related to the lack of direct visual information provided when models receive Chinese characters. For example, the PIXEL model transforms Chinese characters into pixel representations, thereby capturing visual information that significantly enhances performance, as evidenced by its outstanding F1 score of 85. The majority of LLMs handle Chinese text

When decompose Chinese character into its constituent components, you should list its components in the following specific order based on its structure:
 For vertical structures: top to bottom,
 For horizontal structures left to right,
 For wrapping structures: from outside to inside,
 For inlays: main component first, followed by embedded components.
Let's think step by step. First identify the radical of the character. The radical is usually associate 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.
 The components of Chinese character {character} is:

Figure 7: Prompt Line of Component Task

1) **Recognize Basic Components:**
Break down the character into its basic components or radicals. This can help in counting the strokes more accurately.

2) **Count Strokes in Each Component:**
For each component or radical, count the number of strokes. Use the general rules for stroke order to ensure no strokes are missed.

3) **Sum the Strokes:**
Add the stroke counts of all components to get the total stroke count for the character.

4) **Verify the Stroke Count:**
Cross-check the total stroke count with reliable sources or stroke count databases to ensure accuracy.
Output the number of strokes required to write Chinese character {character}:

Figure 8: Prompt Line of Stroke Number Task

In Chinese calligraphy, characters are composed of five standard strokes:

横 (一): This includes 横 and 提
 竖 (丨): This includes 竖 and 竖钩
 撇 (丿): This includes 撇
 捺 (丶): This includes 捺 and 点
 折 (𠃉): This includes all types of 折 such as 横折 and 横钩

1: Recognize the basic components.
Break down the character into its basic components or radicals, as this can help in understanding the structure and stroke order.

2: Apply general stroke order rules.
Recall the general rules for Chinese stroke order:
 Top to bottom
 Left to right
 Outside before inside
 Main before inlays

3: Determine the specific stroke order.
Using the rules and components identified, determine the specific stroke order for the character.

What are the strokes of the Chinese character {character} in order?""

Figure 9: Prompt Line of Strokes Task

Encoding	Structure Acc
Unicode	39.80
Stroke	43.80
Pin Yin	13.85
WuBi	11.81
CangJie	11.66

Table 7: GPT-3.5 Fine-tuning’ Performance on different way of encoding.

similarly to English text: before tokenization, each Chinese character is treated as a unique Unicode character. To further investigate the knowledge gained through this process, we conducted a side experiment on different encoding systems.

Setup. We fine-tuned GPT-3.5 by explicitly switching all Chinese characters in the training and testing documents to various encodings—namely, Unicode, stroke, Pinyin⁸, Wubi, and Cangjie⁹—and evaluated them on the structure recognition task to assess the impact of these representations on the model’s learning ability with visual knowledge of Chinese characters.

Results. The results shown in Table 7 indicate that Unicode encoding performs comparably to stroke encoding and significantly outperforms Pinyin encoding. Although Pinyin is the most common typeface, it is limited to phonetic information. After further investigation, we discovered that the order of Chinese characters in Unicode is highly related to the stroke count and structure of the characters: Unicode is ordered by the stroke count of their indexing radical and the stroke count of remaining parts. However, the full potential of Unicode is diminished by numerous exceptions and a broad spectrum of extensions that complicate its utility in conveying visual knowledge.

where similar structures are likely grouped together with stroke counts in incremental order, as detailed in Figure 6.

⁸Pinyin is the Romanization of the Chinese characters based on their pronunciation. In Mandarin, it’s the standard method for typing Chinese characters.

⁹Wubi and Cangjie are two glyph-based input methods that are uncommon to use.

Input character: 噲

Image: 噲

Response by Ernie-4V:

图片中展示的是一个人手持手机的黑白照片，但根据图像内容描述，并没有提及任何汉字元素。因此，无法直接描述图片中的汉字由哪些部分组成以及它们的结构顺序。然而，图像专家模型答案提到了一个汉字“噲”，但这与图片内容描述不符，因为图片中并没有显示这个汉字。

综上所述，由于图片中并未包含汉字元素，所以无法根据图片内容描述来回答用户关于汉字组成部分和结构顺序的指令。如果确实需要了解某个汉字的组成和结构顺序，建议直接提供该汉字的信息，而不是依赖与汉字无关的图片内容。

The picture shows a black-and-white photo of a person holding a cell phone, but according to the image content description, no Chinese character elements are mentioned. Therefore, it is impossible to directly describe what parts the Chinese characters in the picture are composed of and their structural order. However, the image expert model answer mentioned a Chinese character "噲", but this does not match the picture content description because the Chinese character is not shown in the picture. In summary, since the picture does not contain Chinese character elements, it is impossible to answer the user's instructions about the components and structural order of Chinese characters based on the picture content description. If you really need to know the composition and structural order of a Chinese character, it is recommended to directly provide information about the Chinese character instead of relying on picture content that is not related to the Chinese character.

Figure 10: Example of Ernie-4V response to rare character with English translation.

Input character: 拔

Image: 拔

Response by Ernie-4V:

经过仔细查看，图片中展示的中文字符是“拔”而非“拔”。这个字符“拔”的笔画顺序是：横、竖钩、提、撇、横横、竖。如您仍有疑问，建议查阅现代汉语词典了解更多字符的笔画顺序。

After careful inspection, the Chinese character shown in the picture is "拔" instead of "拔". The stroke order of this character "拔" is: 横、竖钩、提、撇、横横、竖. If you still have questions, it is recommended to consult the Modern Chinese Dictionary to learn more about the stroke order of characters.

Figure 11: Example of Ernie-4V response to extremely similar character with English translation.

Input character: 𠃉

Image: 𠃉


Response by Ernie-4V:

图片中的汉字是**'𠃉'，它的结构是'单一结构'**。

The Chinese character in the picture is **'𠃉', and its structure is 'single structure'**.

Figure 12: Example of Ernie-4V response to part of the character as answer with English translation.

Input character: 𪗇

Image: 

Response by Ernie-4V:

经过深入分析图片中的中文字符，可以确定该字符是**“𪗇”，其结构为左右结构**。。

After in-depth analysis of the Chinese characters in the picture, it can be determined that the character is **"𪗇", and its structure is a left-right structure**.

Figure 13: Example of Ernie-4V response a character with different component part as answer with **English translation**.



Figure 14: Example of Kimi-V reject rarely used character with **English translation**.

D Detailed Radical Prompting Result

D.1 Quantitative Analysis on POS tagging Accuracy

We provide a case analysis for POS tagging in Table 8.

Category	Baseline	RP (Oracle)
Correct& utilize Radical	-	81.2(+81.2)
Correct without	608.6	611.2(+2.6)
Incorrect & utilize Radical	-	41.8(+41.8)
Incorrect without	391.4	265.8 (-125.6)

Table 8: 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. Improvement is shown in green.

D.2 Window size’s impact on POS tagging

We evaluate the impact of different window size in POS tagging with GPT-3.5-Turbo in Table 9.

Window Size	Part-Of-Speech Tagging		
	GPT-3.5-Turbo with GSD		
	B	RP	RP (Oracle)
5	59.08	64.62(+5.5)	67.56(+8.5)
7	60.17	66.55(+6.38)	66.73(+6.56)
9	60.38	67.03(+6.65)	67.23(+6.85)

Table 9: Model performance for POS tagging with different word span sizes

D.3 Radical Prompting Prompts

We provide our prompting lines for POS tagging, NER, and CWS tasks in Figure 15, 16, and 17, respectively.

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. We note that the use of rare Chinese words triggered some safety mechanisms in models such as Gemini-1.5. However, our intent complies fully with the

ethical guidelines and usage policies provided by the API providers.

E.2 Computational Experiments Cost

In our research, we utilized vLLMs for evaluation on Yi 6B, Mistral 7B, Baichuan 13B, and Qwen 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: conllevl.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 word polishing.

```

## Task
Analyze the part of speech (POS) tag of the central word (enclosed in brackets [ ]) in a given section of a sentence with additional information on the component of the Chinese word. The label should be chosen from the following set: {'ADJ', 'PUNCT', 'PRON', 'CCONJ', 'NUM', 'DET', 'X', 'PROPN', 'SCONJ', 'SYM', 'VERB', 'AUX', 'NOUN', 'ADP', 'PART', 'ADV'}

Please note:
1. Label only the center word (the 3rd word) in the 5-word span provided.
2. You should choose only from the label set provided above.
3. Consider the broader spectrum of meanings and functions that a word can embody. For instance, the word "活动" at first glance may seem like a verb meaning "to move" or "to exercise." However, it can also function as a noun, referring to "an activity" or "an event."
4. The complexity of a character—determined by the number of components or the intricacy of each component—can influence its typical POS tag. Words with greater complexity tend to be nouns or pronouns, indicating specific entities or subjects. In contrast, words that are simpler or consist of a single component are more likely to be classified as particles (PART), coordinating conjunctions (CCONJ), or subordinating conjunctions (SCONJ). This pattern reflects the varying linguistic functions these words serve based on their structural complexity.
5. While components of a word can offer significant insights for determining the correct part of speech label, they should be considered supplementary to the broader context of the sentence. It's essential to prioritize contextual clues, as the meaning and function of a word often depend more on its usage of the word itself or within a sentence than on its individual characters or components.

Read the provided sentence carefully and identify the label.
Step 1. Identify the meaning of the center word without using component information. If the meaning is clear, ignore step 2 and go to step 3 without using component information.
Step 2. If the word's meaning is unclear, examine its components to infer potential meanings.
Step 3. Without looking at the context, consider all possible grammatical functions of the word, such as "活动" being both a verb and a noun.
Step 4. Use the sentence's context to determine the most suitable part of speech for the center word.
Step 5. Finalize your choice in JSON format, where the key must be "label" and the value must be the label you have chosen.

## The Provided Sentence
Sentence to Analyze: "{text}" The Component of The Center Word: "{components}" ""

```

```

## Examples
*Example 1 Sentence to Analyze:
"南部， [楠溪] 江下游"
The Component of The Center Word:
"木. 南; ; 溪"
Thought:
1. The meaning of the center word "楠溪" is unclear without additional information.
2. The component information of "楠溪" is "木. 南; ; 溪." This composition suggests that "楠溪" could potentially mean a geographical name as "木" suggesting trees and ";" indicating water.
3. Without consider the context, "楠溪" is most likely to be PROPN but there is a chance that it is NOUN.
4. Considering the sentence's context, "楠溪" is preceded by comma, " " which does not provide useful information, but it is followed by "江", river. Thus, "楠溪" is most likely a proper noun here as the name of the river.
5. Therefore, the most suitable part of speech tag for the center word "楠溪" is PROPN.
Final Answer:
{"label": "PROPN"}

*Example 2 Sentence to Analyze:
"他希望 [能] 共享办公"
The Component of The Center Word:
"么. 月. 七. 匕"
Thought:
1. The meaning of the center word "能" is clear without additional information. "能" means "be able to" in Chinese.
2. Without consider the context, The possible labels for "能" is VERB (when it means "to be able to" or "can") or AUX (when "能" is used to express capability, possibility, or permission).
3. Considering the sentence's context, The sentence structure and the presence of another verb "共享" (to share) immediately after "能" suggest that "能" is serving an auxiliary function rather than acting as a main verb on its own. The speaker's intent is to express a wish or hope, which is a modal use, supporting the use of "能" as an auxiliary (AUX) verb here.
4. Therefore, the most suitable part of speech tag for the center word "能" is AUX.
Final Answer:
{"label": "AUX"}

*Example 3 Sentence to Analyze:
"技术设计 [审查]， 随即"
The Component of The Center Word:
"么. 申. 木. 旦"
Thought:
1. The meaning of the center word "审查" is clear without additional information. "审查" can refer to the process of examining or the action of reviewing documents to ensure they meet certain standards or criteria.
2. Without consider the sentence's context, "审查" can be VERB or NOUN. When "审查" means process of examining, it functions as a verb (VERB). When refer to the action of reviewing documents, it functions as a noun (NOUN).
3. Considering the sentence's context, the structure of the fragment suggests that "审查" is part of a nominal phrase "技术设计审查" (technical design review), indicating a process or event rather than an action being performed at the moment described. "审查" is followed by comma, "，", which does not provide useful information.
4. Therefore, the most suitable part of speech tag for the center word "审查" is NOUN.
Final Answer:
{"label": "NOUN"}

```

Figure 15: Prompt Line of POS tagging.

```

## Task
Perform name entity recognition (NER) task for the sentence below. The label should be chosen from the following set: {'B-LOC', 'B-ORG', 'B-PER', 'I-LOC', 'I-ORG', 'I-PER', 'O'}

Please note:
1. You should choose only from the label set provided above.
2. Strictly tag EACH character in the sentence with BIO standard.
3. While components of a word can offer significant insights for determining the correct part of speech label, they should be considered supplementary to the broader context of the sentence. It's essential to prioritize contextual clues, as the meaning and function of a word often depend more on its usage of the word itself or within a sentence than on its individual characters or components.

Read the provided sentence carefully and identify labels by strictly output steps below.

Step 1. Identify the meaning of the sentence as a whole.

Step 2. Recognize the all entities in the sentence.

Step 3: If an entity has an ambiguous meaning, examine the associated component information as it may offers additional insights.

Step 4: List entities' tags with reasons.

Step 5. Finalize the answer in a JSON format: there should be only one key "label" and the value must be the tag for each character and its corresponding tag combined by "|".

## The Provided Sentence
Sentence to Analyze:
"{text}"

Thought:

```

```

## Examples
*Example 1 Sentence to Analyze:
"在陕西渭南市， 杜雪田被中智选为下一任总监。"
Thought:
Step 1. The sentence situates an event (meeting someone) in a specific location (Weinan City, Shanxi), and describes Du Xuetian being chosen as the next general manager by Zhong Zhi.
Step 2. The entities in this sentence includes: "陕西", "渭南市", "杜雪田", "中智", and "总监"
Step 3. The meaning of "杜雪田" might need a further check: The component of "杜雪田" is "木. 土. 田; 田; 田;" "雨" suggests rain, which might be part of the given female's name to imply purity or natural beauty, while other component does not provide useful information, "杜雪田" is indeed a person name.
The meaning of "中智" might need a further check: The component of "中智" is "中; 知, 日;" "中" suggest central, "知" suggest knowledge, "中智" could suggest an organization that focus on central intelligence or wisdom.
Step 4. List of name entities in this sentence: "陕西" should be tagged as LOC because it is a province in China. "渭南市" should be tagged as LOC because it is a city in ShaanXi. "杜雪田" should be tagged as PER because it is a person name. "中智" should be tagged as ORG because it is a specific organization's name. "总监" should be tagged as O because it is a general position.
Step 5. Final Answer:
{"label": "[在|O], [西|B-LOC], [渭|B-LOC], [南|I-LOC], [市|I-LOC], [，|O], [杜|B-PER], [雪|I-PER], [田|I-PER], [被|O], [中|B-ORG], [智|I-ORG], [选|O], [为|O], [下|O], [一|O], [任|O], [总|O], [监|O], [。|O]"}

*Example 3 Sentence to Analyze:
"如今， 古阳关烽燧还在。"
Thought:
Step 1. The sentence describes the beacon of Guyang Pass still exist.
Step 2. The entities in this sentence are "古阳关" and "烽燧."
Step 3. The meaning of "烽燧" may need additional look. The component of "烽燧" is "{火. 彡. 火. 燧}" suggesting a relation to fire or signals. Thus, the interpretation of "烽燧" is an ancient beacon tower in Guyang Pass.
Step 4. List of name entities in this sentence: "古阳关" should be tagged as LOC because it is a historical location's name. "烽燧" should be tagged as O because it is a general object.
Step 5. Final Answer:
{"label": "[如|O], [今|O], [，|O], [古|B-LOC], [阳|I-LOC], [关|I-LOC], [烽|O], [燧|O], [还|O], [在|O], [。|O]"}

```

Figure 16: Prompt Line of NER.

Task
Analyze the sentence and perform Chinese word segmentation (CWS).

Please note:

1. The answer return must be separate by space.
2. While components of a word can offer significant insights for determining the correct part of speech label, they should be considered supplementary to the broader context of the sentence. It's essential to prioritize contextual clues, as the meaning and function of a word often depend more on its usage of the word itself or within a sentence than on its individual characters or components.

Please read the provided sentence carefully and segment it by strictly following the guide outlined below, step by step:

Step 1. Understand the sentence structure and general meaning

Step 2. Identify words that have unclear meanings.

Step 3. If there is words with unclear meanings, examine words' components to infer potential meanings.

Step 4. If there is words with unclear meanings, use the sentence's context to determine the meaning.

Step 5. Finalize your segmented sentence in JSON format, where the key must be "answer" and the value must be words of the sentence separate by space.

The Provided Sentence
Sentence to segment:
"{text}"

Thought:

Examples

*Example 1 Sentence to Analyze:
"上塘镇位于永嘉县南部，楠溪江下游。"
Thought:
Step 1. The general meaning of the sentence is: 上塘 Town is in the southern part of永嘉 County and is downstream of the楠溪 River.
Step 2. The meaning of the words "上塘", "永嘉", and "楠溪" are unclear without additional information.
Step 3. The component information of "上塘" is "上; 土, 唐." The first character, "上", means "above" or "up". The second component combines "土" (earth or soil) with elements of "唐." Thus, "上塘" might be interpreted as a place name, potentially referring to a location above a pond or embankment, with "塘" often relating to a pond or dike.
The component information of "永嘉" is "永; 土, 口, 丿, 一, 加." "永" means eternal or forever, suggesting a sense of timelessness. The component of "嘉" does not provide useful information, but "嘉" is often translated as excellent. "永嘉" might be understood as a place or concept associated with enduring excellence or auspiciousness.
The component information of "楠溪" is "木, 南; 辶, 奚." This composition suggests that "楠溪" could potentially mean a geographical name as '木' suggesting trees and '辶' indicating water.
Step 4. Final Answer:
{"answer": "上塘镇位于永嘉县南部，楠溪江下游。"}
*Example 2 Sentence to Analyze:
"南山截竹为箏篥。"
Thought:
Step 1. The general meaning of the sentence is: In location NanShan, bamboo is cut to make 箏篥.
Step 2. The meaning of "截竹" and "箏篥" is not clear.
Step 3. The component of "截竹" is "{戹, 钅, 竹}." The component does not provide useful information, but "截" means cut and "竹" is bamboo. Thus, two character should be separated.
The component of "箏篥" is "{竹, 毕, 栗}." "竹" suggest bamboo related, while "毕" and "栗" provided only phonetic clue of the word. The term "箏篥" refers to a traditional Chinese musical instrument and it should be one word.
Step 4. Final Answer:
{"answer": "南山截竹为箏篥。"}
*Example 3 Sentence to Analyze:
"如今，古阳关烽燧还在。"
Thought:
Step 1. The sentence describes the beacon of Guyang Pass still exist.
Step 2. The meaning of "烽燧" may need additional look.
Step 3. The component of "烽燧" is "{火, 夆, 火, 遂}." suggesting a relation to fire or signals. Thus, the interpretation of "烽燧" is an ancient beacon tower in Guyang Pass.
Step 4. Final Answer:
{"answer": "如今，古阳关烽燧还在。"}
}

Figure 17: Prompt line for CWS.