# Evaluating Transformer Models and Human Behaviors on Chinese Character Naming

**Xiaomeng Ma**
The Graduate Center, CUNY
New York, USA
xma3@gradcenter.cuny.edu

**Lingyu Gao**
Toyota Technological Institute at
Chicago, Chicago, USA
lygao@ttic.edu

## Abstract

Neural network models have been proposed to explain the grapheme-phoneme mapping process in humans for many alphabet languages. These models not only successfully learned the correspondence of the letter strings and their pronunciation, but also captured human behavior in nonce word naming tasks. How would the neural models perform for a non-alphabet language (e.g., Chinese) unknown character task? How well would the model capture human behavior? In this study, we first collect human speakers' answers on unknown Character naming tasks and then evaluate a set of transformer models by comparing their performance with human behaviors on an unknown Chinese character naming task. We found that the models and humans behaved very similarly, that they had similar accuracy distribution for each character, and had a substantial overlap in answers. In addition, the models' answers are highly correlated with humans' answers. These results suggested that the transformer models can capture humans' character naming behavior well.[1]

## 1 Introduction

Many aspects of language can be characterized as quasi-regular: The relationship between inputs and outputs is systematic but allow many exceptions. Grapheme-phoneme mapping is an example of such quasi-regularity. For example, the letter string '*-ave*' in English is regularly pronounced as /eɪv/ in GAVE, SAVE, with the exception of /æv/ in HAVE. And human speakers can easily grasp both patterns, e.g., in a nonce word naming experiment, most speakers pronounced the word TAVE as /teɪv/, while some pronounced it as /tæv/ (Glushko, 1979).

To explain the grapheme-phoneme mapping process, many models have been proposed, among which the Dual Route Cascaded (DRC) model and the connectionist model are the two most influential yet opposite models. The DRC model (Coltheart et al., 2001; Coltheart, 1978) proposes that the grapheme-phoneme mapping is implemented in two separate routes: a lexical route that directly maps the word's spelling to its pronunciation through a dictionary-like lookup procedure,[2] and a non-lexical route that applies the grapheme-phoneme corresponding 'rules' to convert the letters to their corresponding pronunciation. The implementation of the DRC model requires domain-specific knowledge, such as spelling to sound rules. In contrast, the connectionist model (Seidenberg and McClelland, 1989; Plaut et al., 1996) proposed that a word's pronunciation is generated through a neural network that takes the orthographic representation as the input and outputs the phonological representation, which does not require specific knowledge of grapheme-phoneme correspondence rules. Both models can explain various behaviors in word identification, such as the faster identification of frequent words compared to infrequent ones. Therefore, there is still an ongoing debate about which model better captures the grapheme-phoneme mapping process.

However, most of these models were tested on alphabetic languages (e.g., English and German), and it is still unclear how these models would be generalized to a non-alphabetic language, such as Chinese. The DRC model seems to be unfit for Chinese because there are no regularities in Chinese that can be defined as grapheme-phoneme corresponding rules (Yang et al., 2009). In addition, Coltheart et al. (2001) asserted that ''the

[1] The code and data for this paper can be found at: https://github.com/xiaomeng-ma/Chinese-Character-Naming.

[2] The lexical route is usually applied to sight words (e.g., 'of', 'and') and words that don't follow grapheme-phoneme correspondence rules (e.g., 'colonel').

Transactions of the Association for Computational Linguistics, vol. 11, pp. 755–770, 2023. https://doi.org/10.1162/tacl_a_00573
Action Editor: Micha Elsner. Submission batch: 12/2022; Revision batch: 2/2023; Published 7/2023.
© 2023 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Chinese, Japanese and Korean writing systems are structurally so different from the English writing system, that a model like the DRC model would simply not be applicable.'' (p. 236). Thus the connectionist model is the only candidate. The majority (81%) of Chinese characters are phono-semantic compounds (Li and Kang, 1993), which consist of a phonetic radical that contains pronunciation information (denoted by pinyin),[3] and a semantic radical that contains semantic information.[4] For example, for the character 晴 (<qing2>, 'sunny'), the left side 日 (<ri4>, 'sun') is the semantic radical, and the right side 青 (<qing1>, 'blue') is the phonetic radical. While the phonetic radical does not contain componential information about the pronunciation, e.g., the first part of the phonetic radical does not represent the first phoneme (e.g., consonant)/syllable onset as letter strings, the relationship between the phonetic radical's pinyin and the character's pinyin is also quasi-regular. Ignoring the tonal differences, the character's pinyin can be categorized into 4 types (Fang et al., 1986): *regular*, the same as the phonetic radical's pinyin; *alliterating*, deviating in the syllable final; *rhyming*, deviating in the syllable onset; and *irregular*, varying in both syllable onset and final (see Table 1 for examples). The process to pronounce an unknown character involves two steps, where the first step is to identify the phonetic radical, and the second step is to apply the regularity pattern of the pinyin. However, there are no reliable cues to identify the phonetic radical, and the regularity patterns are quite arbitrary (Yang et al., 2009). How do Chinese speakers name an unknown character, and how well can the neural models capture the Chinese speakers' behaviors?

| | Example characters |
|---|---|
| *regular* | 清, 情, 圊, 晴 – <**qing**> |
| *alliterating* | 倩 <**q**ian>, 錆 <**q**iang> |
| *rhyming* | 精, 靖, 菁 – <j**ing**> |
| *irregular* | 猜 <cai>, 靚 <liang>, 靛 <dian> |

Table 1: Examples of characters with the phonetic radical 青 <qing>, sorted into different regularity types. Syllable onsets and finals are **bold** when they are the same with the phonetic radical.

In our study, we first collected human speakers' answers on unknown character naming, since there is no study investigating how Chinese adults read unknown characters.[5] We then trained a set of sequence-to-sequence transformer models with different settings on 4,281 phono-semantic characters. Neither human speakers nor models can name the unknown characters accurately, but the transformers have a slightly better average accuracy (47.4%) than the human speakers (45.3%). We then evaluated how closely the results of our aggregated transformers matched those of the human participants, in aspects of the variety of answer types and answer overlaps. In general, both the transformers and human speakers are able to identify the phonetic radical correctly and apply all 4 types of regularities to infer the pinyin, and the transformer models show a high correlation with human data in the proportion of each regularity type. In addition, there is a considerable amount of agreement between the answers generated by our models and those given by humans. Our results demonstrate that transformer models can capture human behavior in unknown Chinese character naming well.

## 2 Related Work and Current Study

Skilled Chinese readers make use of the phonetic radicals to name characters (Chen, 1996; Zhou and Marslen-Wilson, 1999; Ding et al., 2004), and previous studies measured how phonetic radicals influence character naming in two ways: regularity and consistency (Fang et al., 1986; Hue, 1992; Hsu et al., 2009). The regularity is exemplified in Table 1, and the consistency is defined as the number of characters that share the same phonetic radicals and pinyin. For example, there are 12 characters sharing the phonetic radical 青 <qing> in Table 1, among which 3 characters (精, 靖, 菁) have the same pinyin <jing>, so the consistency score for these characters is 0.25 (3/12). Many studies have found *regularity and consistency effects* for human speakers—the *regular* and more

[3] Chinese characters use pinyin to represent the pronunciation. The pinyin system consists of 24 syllable initials (mostly contain a consonant), 34 syllable finals (mostly contain a vowel or vowels), and 4 tones.

[4] The phonetic radical and semantic radical are mutually exclusive, and they are defined in the ancient Chinese dictionary《說文解字》*'Shuowen Jiezi'*.

[5] Previous studies have focused on children's behavior on unknown character naming and found that children made errors in identifying the incorrect phonetic radical, as well as applying the incorrect regularity pattern (Lam, 2008, 2014).

consistent characters are named faster and more accurately, and these effects are stronger for low-frequency characters than high-frequency ones (Lien, 1985; Liu et al., 2003; Tsai et al., 2005).

Previous studies of Chinese character modeling with phonetic radicals as inputs have successfully simulated the regularity effect and consistency effect. Yang et al. (2009) trained a feed-forward network on 4,468 Chinese characters and tested the model on 120 characters (seen in the training). The input to the model includes the character's radicals and radicals' positions (e.g., left-right, up-down).[6] The output of the model is the phonological features (e.g., stop, lateral) of the character's pinyin. They also measured the human speakers' response latency[7] on each of the 120 test characters. By comparing the human speakers' response latency and the model's sum squared error, they found very similar regularity and consistency effects. In addition, Hsiao and Shillcock (2004, 2005) trained a feed-forward model on 2,159 left-right structured characters, with each character appearing according to its log token frequency. The input included each character's radicals, and the output was the character's pinyin. They analyzed the training accuracy of the model and found the model's sum squared errors lower for the *regular* characters, which successfully simulated the regularity effect.

The regularity and consistency effect revealed that both human speakers and the neural models utilized the statistic distribution of phonetic radicals in naming familiar characters. However, these effects can not be applied in unknown character naming since the speakers don't know the statistics of these characters. Therefore, we proposed a new metric (saliency of the phonetic radical) to measure how the phonetic radicals influence the speaker's unknown character naming behavior. Saliency is defined as the fraction of the *regular* characters among all characters sharing the same phonetic radical. For example, the phonetic radical 青 <qing> appeared in 12 characters in Table 1, among which 4 characters （清，情，圊，晴） are *regular*. Thus the saliency score of 青 <qing> is 0.33 (4/12). The more salient a phonetic radical is, the more likely the character that contains it is pronounced the same as its pinyin.

We hypothesized that the human speakers would show a saliency effect in unknown character naming - they would name the characters more accurately if the phonetic radical is more salient. We expected to find a similar saliency effect in the models. In addition, we also closely examine the models' answers and humans' answers to investigate if the models can represent the human speaker's behavior.

## 3 Data

The base character dataset consists of 4,341 Chinese characters constructed from the IDS dataset in CHISE project (Morioka, 2008). The original IDS (Ideographic Description Sequence) dataset contains 18,347 characters used in China, Japan, and Korea with the decomposition of each character's phonetic and semantic radicals.[8] The character selection criterion include: 1) is used in Chinese; 2) is a phono-semantic compound; 3) has a left-right structure.[9] The character's pinyin, along with its phonetic and semantic radical's pinyin, was collected using the pinyin package. The frequency of each character was extracted from BLCU Corpus Center (Xun et al., 2016). We further labeled each character's regularity: *regular*, *alliterating*, *rhyming*, and *irregular* as described in Table 1. In addition, we calculated each phonetic radical's saliency.

There are 660 radicals after decomposing the 4,341 characters, among which 46 radicals only serve as the semantic radicals; 493 radicals only serve as the phonetic radicals; 121 radicals serve as both semantic and phonetic radicals. Each radical appears in 7 characters on average, with a range of 1 to 30. Eighty percent of the characters in our database have the phonetic radical on the right, with many exceptions, e.g., the semantic radical '戈' <ge> always appears on the right.

### 3.1 Test Data

We selected 60 characters with different phonetic radicals from the dataset as our test data, which

[6] There are 10 different Chinese character structures clustered by the arrangement of the character radicals, e.g., left-right (日+ 青 = 晴), top-down (相 + 心 = 想), and enclosure (口+或=國). The left-right structure is the most common type (71%) (Hsiao and Shillcock, 2006).

[7] Response latency measures the response speed, usually in milliseconds.

[8] The phonetic and semantic radicals are decomposed according to *Shuowen Jiezi*《說文解字》.

[9] Following Hsiao and Shillcock (2004), we only selected left-right structure to make sure that the character's structure is not a variable in our study.

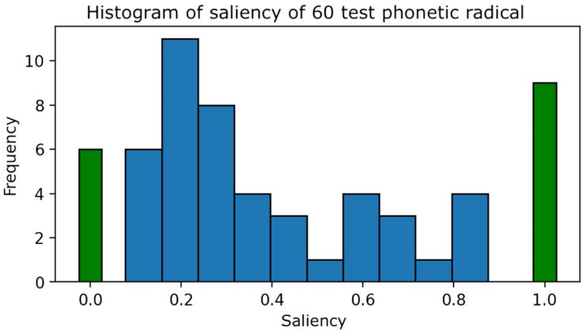

Figure 1: The histogram of saliency scores for 60 test characters' phonetic radicals.

| Regularity Type | # pinyin | # character |
|---|---|---|
| *regular* | 30 | 30 |
| *alliterating* | 6 | 6 |
| *rhyming* | 24 | 19 |
| *irregular* | 28 | 20 |

Table 2: The distribution of regularity types of pinyins and characters for the test data.

are listed in Table 14 in Appendix B. The test characters are selected following two criteria to ensure that human speakers are unfamiliar with the character, while familiar with the phonetic radicals: 1) the character appears less than 5 times in the whole corpus, 2) the phonetic radical in each character appears in more than 4 other characters. The average saliency score for these phonetic radicals is 0.43, with the score distribution shown in Figure 1. Among these characters, 22 of them have more than one pinyin, e.g., '硞' (<que>, <ke>, <ku>), which yields 88 pinyins for 60 characters. The distribution of the regularity type for the test characters is shown in Table 2.

### 3.2 Training Data

We exclude the 60 test characters and use the rest of the characters as our training data (4,281). The *regular* is the most common type (42.7%), followed by *irregular*, *rhyming*, and *alliterating*. Since many of the characters have extremely low frequency and are not known to the Chinese speakers, we used three training datasets with characters of different frequencies to represent the native speakers' vocabulary size. The ALL dataset used all 4,281 characters. The MID dataset consists of 2,140 characters whose frequencies are in the top 50% percentile. The HIGH dataset consists of 1,070 characters with frequencies in the top 25% percentile. The statistics of these training sets are shown in Table 3. Each training set has similar regularity distribution.

| Training Data | ALL | MID | HIGH |
|---|---|---|---|
| # characters | 4,281 | 2,140 | 1,070 |
| *regular* (%) | 42.7 | 43.3 | 42.1 |
| *alliterating* (%) | 7.8 | 8.1 | 8.7 |
| *rhyming* (%) | 23.6 | 23.3 | 22.5 |
| *irregular* (%) | 25.9 | 25.3 | 26.7 |

Table 3: Number of characters and percentage of regularity types for our training datasets.

## 4 Human Experiment

A total of 55 native speakers of Mandarin participated in this study. All of them are able to read and write in traditional Chinese scripts and pinyin. The average age is 26.3 years, and 80% of them have an education background of college or above. In the experiments, they were asked if they knew the character and prompted to type the pinyin of the character. The detailed experiment procedure and sample questions are described in Appendix A.

### 4.1 Results: Human Answer Accuracy

In general, the test characters are unknown to the participants.[10] The accuracy is calculated on the syllable onset and final, ignoring the tone, since tones are more affected by the speaker's accent than syllable onsets and finals. For polyphone characters, as long as the participant named one correct pinyin, we counted it as correct. The average accuracy for all participants is 45.3% (27 out of 60 characters), with a range of 26.7%–68.3%. Some characters are more difficult to name than others. For example, 8 characters' accuracies are 0, meaning that none of the participants named them correctly. The character's accuracy is calculated as the proportion of participants who named it correctly, ranging from 0%–98.2%. There is a strong positive correlation between the character's accuracy and its phonetic radical's saliency

[10] Very few participants indicated that they knew one or two test characters. For those who indicated that they knew the character, they still answered its pinyin incorrectly.

煔 <shan>, <qian>, ‘sparkle’
Phonetic Radical: 占 <zhan>, ‘to seize’
Semantic Radical: 炎 <yan>, ‘fire’

| Answer Type | Answer(s) | $P_p$ (%) |
|---|---|---|
| *regular* | <zhan> | 36.4 |
| *alliterating* | <zhen> | 1.8 |
| *rhyming* | <dan> | 3.6 |
| *irregular* | <nian>, <jian>, <dian>, <pou> <yi>, <tie> | 34.6 |
| *semantic* | <yan> | 23.6 |

Table 4: The answer types and production probability of human answers for polyphone ‘煔’.

(r = 0.62), which confirms our hypothesis about the saliency effect. The more salient the phonetic radical is, the more participants named the character correctly. The accuracy measures how well the human speakers can grasp the grapheme-phoneme distributional patterns in Chinese. The results show that even native speakers can not accurately predict the pronunciation of an unknown character, which reflects the complex nature of Chinese grapheme-phoneme mapping system.

### 4.2 Results: Human Answer Variability

Since the participants named the character’s pinyin differently, each character has a variety of unique answers. On average, each character has 6.7 answers, with a minimum of 2 answers and a maximum of 15 answers. The number of answers is negatively correlated with the saliency of the phonetic radical (r = -0.51), such that the more salient the phonetic radical, the fewer number of answers the speakers guessed.

We defined 5 answer types based on regularity. The participants either guessed the character’s pinyin the same as its phonetic radical’s (*regular*), or changing the syllable final (*alliterating*), the syllable onset (*rhyming*), or both (*irregular*), or mistakenly used the semantic radical to name the character (*semantic*).[11] We presented the answer types for character ‘煔’ as an example in Table 4, and defined the production probability $P_p$ by the proportion of participants named that answer type.

| Answer Type | Average $P_p$ (%) | Range of $P_p$ (%) |
|---|---|---|
| *regular* | 58.0±25.8 | 0–98.2 |
| *alliterating* | 6.8±16.4 | 0–81.8 |
| *rhyming* | 13.0±18.9 | 0–81.8 |
| *irregular* | 20.6±20.0 | 0–72.7 |
| *semantic* | 1.6±4.6 | 0–23.6 |

Table 5: The average production probability and its range for each answer type in human answers.

The average production probability for each type is listed in Table 5. Most of the participants are able to identify the phonetic radical correctly, as the average production probability of the semantic type is only 2%. The regular answer type has the highest production probability (58%), suggesting that the participants are more likely to name the character the same as its phonetic radical. The production probabilities of answer types for each character are plotted in Figure 3 in Section 6.

## 5 Transformer Model

To model the joint probability of the syllable onset and final, we used seq-to-seq transformers (Vaswani et al., 2017) to generate the pinyin of Chinese characters trained from scratch.[12]

### 5.1 Experiment Setup

Both encoder and decoder of all our models had 2 layers, 4 attention heads, 128 expected features in the input, and 256 as the dimension of the feed-forward network model. For training, we split the dataset into train/dev splits of 90/10, and replace those tokens that appear once in training data by ⟨unk⟩. We also set dropout to 0.1, batch size to 16, and used the Adam optimizer (Kingma and Ba, 2015) with varied learning rates in the training process, computed according to Vaswani et al. (2017). We used 5 different random seeds, and trained 40 epochs with early stopping for all

[11] When examining the data, we found that some participants named the character the same as its semantic radical. We loosely defined this type of error as *semantic* type. It could also be that the participants applied *irregular* on the phonetic radical, and the pinyin happened to be the same as the semantic radical’s pinyin. However, there’s no way to confirm this. We asked some of our participants (with linguistic background) to explain how they guessed the pinyin, and none of them could articulate their thinking process.

[12] We did not use a classification model because there are certain rules in pinyin formation (e.g., /ü/ cannot follow /b/, /p/, /m/, /f/), which requires the model to learn the syllable onsets and finals jointly.

of our experiments. For inference, we set beam size to 3.

## 5.2 Experiment 1

We trained a set of models to simulate the grapheme-phoneme mapping process in Chinese speakers. Our BASE model used the phonetic radical's orthographic forms to generate syllable onset and final (without tone) of the target character. We further examined whether identifying the phonetic radical before generating the syllable onset and final would improve the model's performance. We labeled the phonetic radical's position (left or right) with two methods: LABEL$_{m}$ and LABEL$_{s}$. LABEL$_{m}$ used the true position of the phonetic radical as the ground truth label. Besides, since human speakers do not always identify the phonetic radical's position correctly, LABEL$_{s}$ labeled the position of the phonetic radical based on the phonetic similarity. We calculated the phonetic similarity between the character's pinyin and the two radicals' pinyins using the Chinese Phonetic Similarity Estimator (Li et al., 2018). The radical with higher phonetic similarity was labeled as the phonetic radical.[13] We further labeled the regularity type of the characters based on LABEL$_{m}$ and LABEL$_{s}$, hence yielding LABEL$_{mr}$ and LABEL$_{sr}$. Examples of input and gold output in the training data are shown in Table 6. All the models were trained on ALL, MID, and HIGH datasets as described in section 3.2.

Since previous studies suggested that the regularity and consistency effects are more prominent for the characters with low frequency than high frequency (e.g., Ziegler et al., 2000; Chen et al., 2009), the frequency of the known characters might also influence how participants predict the unknown characters. We further added the frequency label as an input feature in the full training data as the ALL+FREQ model. The characters were categorized into four categories based on their frequency: 'rare' (frequency = 1), 'low' (1 < frequency $\leq$ 50% percentile), 'mid' (50% percentile < frequency $\leq$ 75% percentile), and 'high' (frequency > 75% percentile). The distribution of regularity types is similar for the characters with different frequencies. The summary of the number of characters and each regularity type can be found in Appendix B, Table 12.

| Input | Begin, 火, 各, End |
|---|---|
| Model | Output |
| BASE | Begin, l, uo, End |
| LABEL$_{m}$ | Begin, right, l, uo, End |
| LABEL$_{s}$ | Begin, left, l, uo, End |
| LABEL$_{mr}$ | Begin, right, irregular, l, uo, End |
| LABEL$_{sr}$ | Begin, left, rhyming, l, uo, End |
| Condition | Input |
| ALL+FREQ | Begin, 火, 各, high, End |
| Condition | Output (BASE model as an example) |
| [+Shuffle] | Begin, uo, l, End |
| [+Tone] | Begin, l, uo, 4, End |

Table 6: Input and gold output in the training data of our models and conditions for character '烙'<luo4>, tokens are separated by comma.

In addition, we added two conditions for output in training all models: Shuffling and Adding tones. We shuffle the position of the syllable onset and final in model output to explore the impact of the generated order since we don't know if the human speakers identify the syllable onset or syllable final first in character naming. We also add tones before the 'End' token in the generation to see whether it improves the model performance. Examples of input and output of the conditions are shown in Table 6. In total, there are 80 types of models with different settings.

**Accuracy Results** We calculated the test accuracy the same way as for the human data: We only counted the accuracy of the syllable onset and final. For polyphone characters, as long as the model predicted one correct pinyin, it is counted as correct. The average accuracy of all 400 models (80 types x 5 random seeds) is 42.1%, which is significantly lower than the humans' accuracy (45.3%, t = 3.15, p<0.01). The average accuracy of each type of model is listed in Table 7. The best performing model is ALL+FREQ with LABEL$_{m}$ without tone and with shuffling, which achieved an accuracy of 50.3%. Compared to the BASE model, adding the label of phonetic position label and the character's regularity label usually could improve the model's accuracy. Adding tone would

[13] For example, the character '烙' <luo4> ('flatiron') consists of the semantic radical '火' <huo3> ('fire') and the phonetic radical '各' <ge4> ('each'). The distance between <luo4> and <huo3> is 7.5, and the distance between <luo4> and <ge4> is 35.6. For LABEL$_{s}$, the output radical should be 'left', although the left radical '火' is the semantic radical.

| data | label | –T–S | –T+S | +T–S | +T+S |
|---|---|---|---|---|---|
| ALL | BASE | 49.3 | 49.3 | 42.3 | 46.0 |
| | LABEL$_{m}$ | 48.0 | 49.7 | 45.3 | 47.7 |
| | LABEL$_{s}$ | 46.0 | 45.3 | 42.3 | 48.7 |
| | LABEL$_{mr}$ | 47.0 | 48.7 | 48.7 | 49.7 |
| | LABEL$_{sr}$ | 44.0 | 47.3 | 45.0 | 48.3 |
| MID | BASE | 41.7 | 41.3 | 38.7 | 41.7 |
| | LABEL$_{m}$ | 44.3 | 43.0 | 44.0 | 42.3 |
| | LABEL$_{s}$ | 41.3 | 43.3 | 41.3 | 42.3 |
| | LABEL$_{mr}$ | 42.0 | 40.3 | 39.7 | 44.3 |
| | LABEL$_{sr}$ | 37.7 | 42.7 | 39.0 | 42.0 |
| HIGH | BASE | 28.7 | 32.3 | 29.3 | 32.3 |
| | LABEL$_{m}$ | 36.3 | 34.7 | 30.7 | 35.3 |
| | LABEL$_{s}$ | 32.7 | 36.0 | 30.0 | 34.0 |
| | LABEL$_{mr}$ | 31.3 | 31.7 | 31.3 | 32.0 |
| | LABEL$_{sr}$ | 32.3 | 32.0 | 31.0 | 33.7 |
| ALL+ FREQ | BASE | 46.7 | 47.0 | 47.7 | 46.7 |
| | LABEL$_{m}$ | 49.7 | 50.3 | 47.3 | 47.0 |
| | LABEL$_{s}$ | 45.3 | 47.3 | 47.0 | 48.3 |
| | LABEL$_{mr}$ | 46.3 | 49.3 | 47.0 | 48.0 |
| | LABEL$_{sr}$ | 47.7 | 44.7 | 44.0 | 47.7 |

Table 7: The average accuracy (over 5 seeds) on the test set for models trained on HIGH, MID, or adding frequency label as input features on ALL. +T, –T, +S, –S refers to adding tone, no tone, shuffling, and no shuffling, respectively.

generally hurt the model's accuracy. Shuffling the syllable onset and final and adding the frequency label in the input would not change the model's accuracy.

### 5.3 Experiment 2

In Experiment 1, the input of our models only used the orthographic form of the radicals, which is how the previous literature described the Chinese grapheme-phoneme mapping process. However, the models might not have enough data to learn the full mapping from radicals to pinyin because many radicals only appeared once or twice in the training data since we only included compound characters with the left-right structure. For example, the phonetic radical '乘' <cheng> only occurred once in the character '剩' <sheng> in the training data.[14] The models would not be able to accurately learn the pinyins of these radicals. However, human speakers know the pinyin of most radicals, since many radicals are also commonly used as stand-alone characters, e.g., '乘' is a stand-alone character meaning 'to multiply'. In order to better model the human speakers, it is necessary to inject pinyin of the radicals as external information to the model. The model would also benefit from the added radicals' pinyin to generate the character's pinyin.

| Input | Begin, 火, h, uo, 3, End, 各, g, e, 4, End |
|---|---|

Table 8: Input in the training data for Experiment 2 using '烙' <luo4> as an example.

In addition, pinyin also plays an important role in modern Chinese speakers' reading and spelling experience. Pinyin is a Romanized phonetic coding system created in 1958 to promote literacy (Zhou, 1958). In the information age, pinyin has become indispensable in Chinese speakers' lives because it is the dominant typing system for computers, smartphones, and electronic devices. The prevalent experience of typing characters through pinyin has challenged the traditional view that Chinese characters are processed purely through orthographic forms (Tan et al., 2013). Many recent studies have found that pinyin mediates the character recognition process (Chen et al., 2017; Lyu et al., 2021; Yuan et al., 2022). To better capture modern Chinese speakers' character naming process, it is necessary to incorporate the radical's orthographic form as well as its pinyin in our models.

Therefore, in Experiment 2, we added the radical's pinyin (syllable onset, syllable final, and tone) in the input, as shown in Table 8. We used the same model variations as in Experiment 1[15] and trained 80 different types of models (5 random seeds for each type) with the new input. The training settings are the same as Experiment 1.

**Accuracy Results** Adding pinyin to the input has increased the model's accuracy.[16] The average accuracy of 400 models in Experiment 2 is 47.4%, which is significantly higher than the

[14] We choose the first pinyin from the pinyin package for polyphone radicals.

[15] For the output, we added LABEL$_{m}$, LABEL$_{s}$, LABEL$_{mr}$, LABEL$_{sr}$ as well as adding tone and shuffling. For the input, we added frequency label to create ALL+FREQ.

[16] We cannot fully rule out the possibility that the increased accuracy is due to the model having longer inputs with pinyin instead of the model making use of the phonetic information. However, the input length might not have a significant impact on the models because our models with frequency labels (ALL VS ALL+FREQ) also vary in input lengths but the accuracies didn't change much.

humans' accuracy (t = −2.7, p <0.01). The accuracy for each type of model is listed in Table 11 in Appendix B. The best performing model is ALL+FREQ with LABEL$_{mr}$ without tone and with shuffling, which achieved an accuracy of 55%. The effects of different labels, adding tone, and shuffling are similar to the models in Experiment 1.

## 6 Comparison Between Models' Results and Human Behaviors

In this section, we compared transformer models' results in Experiments 1 (MODEL[−PINYIN]) and 2 (MODEL[+PINYIN]) with human performance. Since human participants are different, i.e., they have different vocabularies, and they may use different strategies to identify the phonetic radical, we used all 80 models in each experiment to represent the human variety. Following Corkery et al. (2019), each random initialization was also treated as an individual participant. Therefore, the sample size for the human participants is 55, and the sample size for the models in each experiment is 400 (80 models × 5 initializations). We focused on three types of similarities: 1) accuracy, i.e., do humans and models show similar accuracy on each character? 2) overlap, i.e., do humans and models predict the same pinyin for each character? 3) variability, i.e., do humans and models have similar answer regularity patterns?

**Accuracy** We calculated each character's accuracy for MODEL[−PINYIN] and MODEL[+PINYIN]. First, both models showed saliency effect: The model's character accuracy is positively correlated with saliency score (Pearson $r$ = 0.48 for MODEL[−PINYIN] and $r$ = 0.57 for MODEL[+PINYIN]), which is not significantly different from humans' saliency correlation ($r$ = 0.62). In addition, there's a strong correlation between human character accuracy and both models' character accuracy (MODEL[−PINYIN] $r$ = 0.79, MODEL[+PINYIN] $r$= 0.88), suggesting that the humans and models are in high agreement. In conclusion, the transformer models' answers are very similar to the human answers in terms of character accuracy.

**Overlap** The overlap rate was computed to measure to what extent different human speakers (and models) predict the same answers for each character. For example, if participant 1 and 2 have 30 same answers, then the overlap rate = 50% (30/60). Among 1,485 answer pairs of 55 human speakers, the average overlap rate of human-human is 50.2%, with a range of 25.0% – 73.3%. For MODEL[−PINYIN], among 400 models and 55 speakers, the average overlap rate for 22,000 answer pairs is 39.6%, with a range of 12.0% – 66.7%. For MODEL[+PINYIN], the average overlap rate of 22,000 answer pairs is 45.2%, with a range of 16.7% – 71.7%. Both models' overlap rates are significantly lower than the human-human overlap rate, and MODEL[+PINYIN]'s overlap rate is significantly higher than MODEL[−PINYIN]. In addition, we computed the human-model overlap rate for different models, with 275 answer pairs for each model (5 random seeds × 55 human speakers).[17] The best model for MODEL[−PINYIN] is ALL with LABEL$_{mr}$ with tone and without shuffling, with an overlap rate of 45.6%. The best model for MODEL[+PINYIN] is ALL+FREQ with LABEL$_{s}$ with tone and without shuffling, with an overlap rate of 50.1%, which is not significantly different from human-human overlap rate. The overlap results are summarized in Table 9. The density plot of the overlap rate for human-human, human-all models, and human-best model is shown in Figure 2. In general, the humans' answers are more similar to each other than to the models' answers. MODEL[+PINYIN]'s answers are more similar to human answers than MODEL[−PINYIN].

| | Overlap rate | Range |
|---|---|---|
| Human - Human | 50.2±7.0 | 25.0–73.3 |
| Transformer models - Human | | |
| All MODEL[−PINYIN] | 39.6*±7.6 | 11.7–66.7 |
| Best MODEL[−PINYIN] | 45.6*±5.8 | 28.3–61.7 |
| All MODEL[+PINYIN] | 45.3*±7.0 | 16.7–71.7 |
| Best MODEL[+PINYIN] | 50.1±6.1 | 31.7–66.7 |

*indicates significantly smaller than 50.2.

Table 9: The average overlap rate (%) and its range for human-human and transformer-human comparison.

**Variability** Like human speakers, transformer models also produce different answers for each character. We categorized these answers based

[17] See the detailed overlap results for MODEL[−PINYIN] and MODEL[+PINYIN] in Table 13, Appendix B.

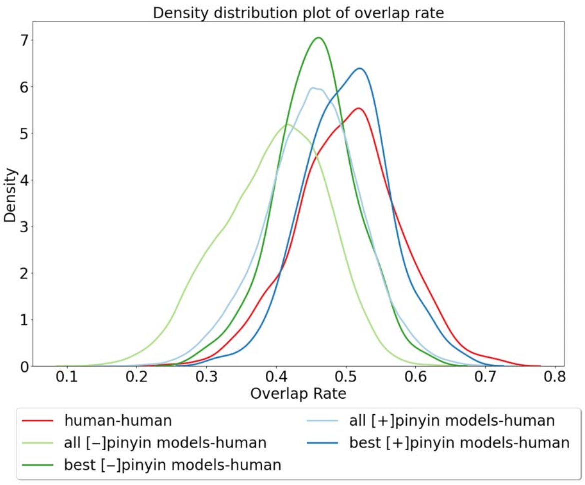

Figure 2: Density plot of the overlap rate.

| | MODEL [−PINYIN] | Cor. | MODEL [+PINYIN] | Cor. |
|---|---|---|---|---|
| *Reg.* | 39.4*±32.6 | $\rho$ 0.72<br>$r$ 0.71 | 52.9±30.3 | $\rho$ 0.70<br>$r$ 0.72 |
| *Alli.* | 10.9±21.4 | $\rho$ 0.59<br>$r$ 0.95 | 10.1±19.7 | $\rho$ 0.51<br>$r$ 0.85 |
| *Rhym.* | 22.6*±28.4 | $\rho$ 0.64<br>$r$ 0.58 | 20.6±25.1 | $\rho$ 0.70<br>$r$ 0.62 |
| *Irr.* | 26.6±28.1 | $\rho$ 0.55<br>$r$ 0.50 | 16.1±20.6 | $\rho$ 0.67<br>$r$ 0.58 |
| *Sem.* | 0.5*±1.7 | $\rho$ 0.39<br>$r$ 0.30 | 0.2*±0.9 | $\rho$ NA†<br>$r$ 0.09 |

* indicates significantly different from human ($P_p$).
NA† is due to too many zeros in the data that the correlation cannot be calculated.

Table 10: The average production probability ($P_p$) of each answer type and their correlation ($\rho$ and $r$) with humans for MODEL[−PINYIN] and MODEL[+PINYIN].

on their regularity type and calculated the models' averaged production probability ($P_p$) for each answer type, as listed in Table 10. We further calculated Spearman correlation ($\rho$) and Pearson correlation ($r$) between the production probability of each type in human answers and the models' answers on each character (N = 60). All the regularity types are highly correlated except for the *semantic* type. The models did not produce as many *semantic* type answers as humans, suggesting that the models are better at identifying the phonetic radical than humans. In addition, we also calculated the cross-entropy between the humans and the models on the production probability of 5 regularity types. The cross-entropy for MODEL[−PINYIN] is H(human, MODEL[−PINYIN]) = 1.79 and for MODEL[+PINYIN] is H(human, MODEL[+PINYIN]) = 1.74, suggesting that MODEL[+PINYIN] is slightly more similar to the human results than the MODEL[−PINYIN].

The production probability of different regularity types for each character is shown in Figure 3. The answer type patterns are very similar for humans and models except for the *semantic* type. Humans produced *semantic* type answers for 15 characters, while both our models produced *semantic* type for fewer characters with a much smaller production probability. This implied that phonetic radicals are identified differently by humans and transformer models. Humans are affected by a wide range of linguistic knowledge in identifying the phonetic radical, including the semantic meaning of the radical, vocabulary size, and reading comprehension (Anderson et al., 2013; Yeh et al., 2017). The models did not receive these extra inputs, and thus did not closely capture human behavior on the *semantic* answer type.

## 7 Conclusion and Discussion

**Conclusion** We evaluated transformer models and human behaviors on an unknown Chinese naming task. This task is difficult for both humans and transformer models, as the average accuracy is lower than 50%. Humans have higher accuracy than MODEL[−PINYIN] and lower accuracy than MODEL[+PINYIN], and the models and the humans have very similar performances. First, saliency effects were found in both human data and the models' results, suggesting that both models and humans utilize the statistical distribution of the phonetic radical to infer the character's pinyin. Further, although humans' answers are more similar to each other, our models also achieved a substantial overlap with humans' answers. Additionally, the production probability of each answer type is highly correlated between models and humans (except for *semantic* type), suggesting that both models and humans are able to apply all regularity patterns in producing answers. Finally, models with radical's pinyins in the input are more similar to humans and achieved higher accuracy.

**Capturing Quasi-regularity** Our work is also related to the long-standing criticism that the

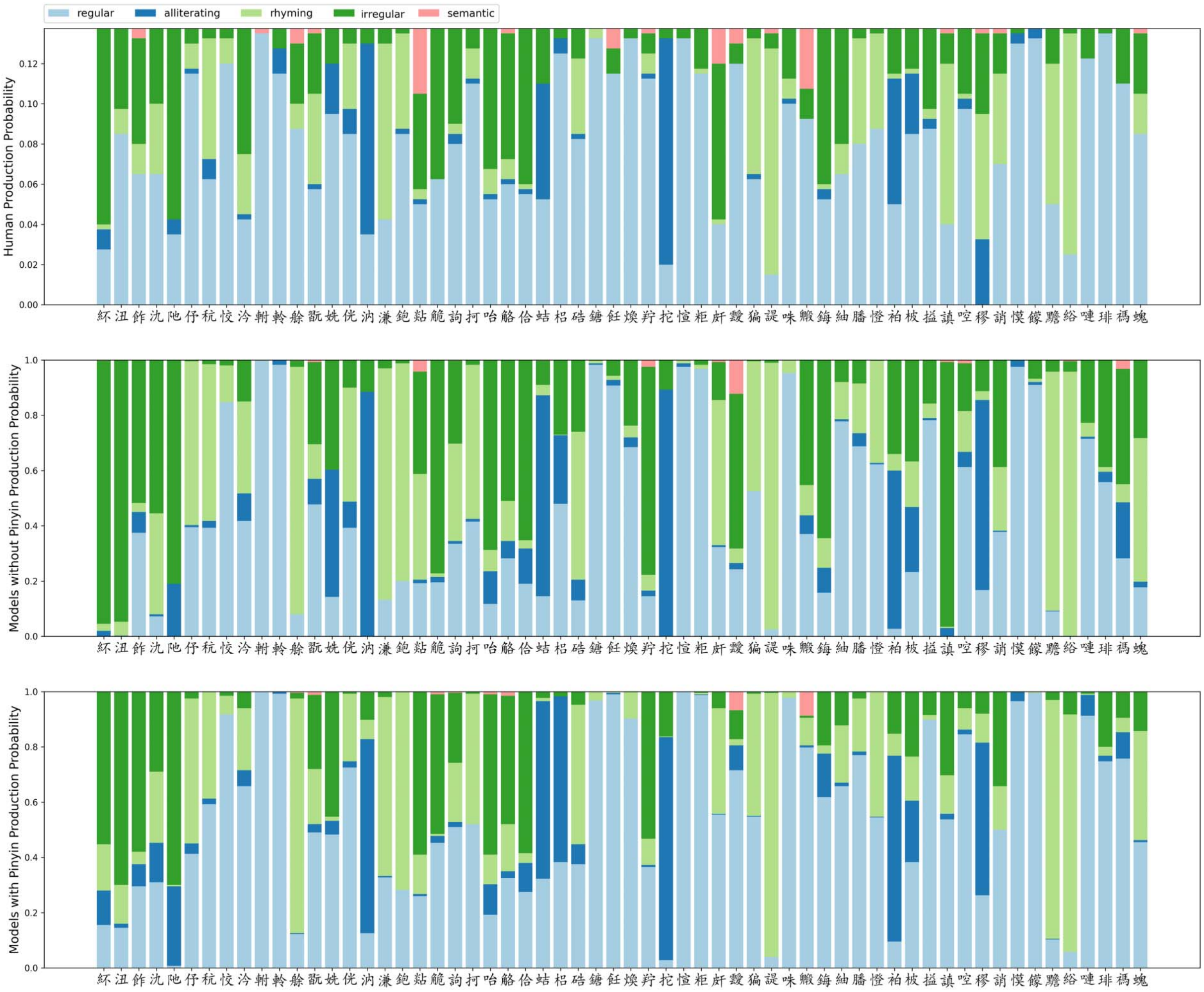

Figure 3: The production probability of 5 answer types produced by humans (top), MODEL[–PINYIN] (middle), and MODEL[+PINYIN] (bottom).

neural networks may only learn the most *frequent* class and can not extend other minority classes, thus would fail to learn the quasi-regularity in languages (Marcus et al., 1995). Previous studies on morphological inflections have shown that the neural models overgeneralized the most frequent inflections on nonce words and had almost no correlation with humans' production probability on the less frequent inflections (e.g., $\rho$ = 0.05 for the /-er/ suffix in German plural [McCurdy et al., 2020], and $r$ = 0.17 for irregular English verbs [Corkery et al., 2019]). However, our results showed that the transformer models could learn the quasi-regularity in Chinese character naming, that the models produce all answer types, and the production probability of each type is highly correlated with human data.

However, our results do not contradict the previous studies. Chinese character naming and morphological inflection both exhibit quasi-regularity, but the two domains are very different: The patterns in Chinese character naming are less rule-governed. This paper's contribution to the debate of quasi-regularity in language processing is not to provide a 'yes' or 'no' answer; instead, we used a novel task and showed that the neural models have the potential to model human behaviors in learning quasi-regularity. We hope our study could inspire future work in this field to apply diverse tasks and conduct more detailed examinations of neural models' ability in learning quasi-regularity.

**Modeling Chinese Reading with Neural Network** Our study also contributed to the current debate of whether reading skill is acquired by a domain-general statistical learning mechanism (Plaut, 2005), or language-specific knowledge such as the DRC model (Coltheart et al., 2001). Our results demonstrated that a general statistical learning mechanism (implemented as the transformer model) could learn Chinese grapheme-phoneme mapping. We not only successfully simulated the general saliency effects in humans' unknown character naming behavior, but also showed in details that the answers produced by models and humans are highly similar. Another contribution to modeling Chinese reading is that we are the first study that incorporated the radicals' pinyin in the model. Models with pinyin as input not only had better accuracy, but also are more similar to human behavior. Our results echoed the recent literature on the pinyin effect. For modern Chinese speakers who type characters through pinyin more often than hand-writing characters, pinyin can be an important mediator for the grapheme-phoneme mapping process.

## Acknowledgments

We would like to thank the anonymous reviewers and action editor for their valuable feedback. We would also like to thank Kevin Gimpel, Allyson Ettinger, Virginia Valian, Martin Chodorow, and Kyle Gorman for their insightful discussions and suggestions.

## References

Richard C. Anderson, Yu-Min Ku, Wenling Li, Xi Chen, Xinchun Wu, and Hua Shu. 2013. Learning to see the patterns in Chinese characters. *Scientific Studies of Reading*, 17(1):41–56.

Hsin-Chin Chen, Jyotsna Vaid, and Jei-Tun Wu. 2009. Homophone density and phonological frequency in Chinese word recognition. *Language and Cognitive Processes*, 24(7–8):967–982.

Jingjun Chen, Rong Luo, and Huashan Liu. 2017. The effect of pinyin input experience on the link between semantic and phonology of chinese character in digital writing. *Journal of Psycholinguistic Research*, 46(4):923–934.

Yi-Ping Chen. 1996. What are the functional orthographic units in chinese word recognition: The stroke or the stroke pattern? *The Quarterly Journal of Experimental Psychology: Section A*, 49(4):1024–1043. https://doi.org/10.1080/713755668

Max Coltheart. 1978. Lexical access in simple reading tasks. *Strategies of Information Processing*, pages 151–216.

Max Coltheart, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108(1):204. https://doi.org/10.1037/0033-295X.108.1.204, PubMed: 11212628

Maria Corkery, Yevgen Matusevych, and Sharon Goldwater. 2019. Are we there yet? Encoder-decoder neural networks as cognitive models of english past tense inflection. In *57th Annual Meeting of the Association for Computational Linguistics*, pages 3868–3877. Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/P19-1376

Guosheng Ding, Danling Peng, and Marcus Taft. 2004. The nature of the mental representation of radicals in chinese: a priming study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(2):530. https://doi.org/10.1037/0278-7393.30.2.530, PubMed: 14979822

Sheng-Ping Fang, Ruey-Yun Horng, and Ovid J. L. Tzeng. 1986. Consistency effects in the Chinese character and pseudo-character naming tasks. *Linguistics, Psychology, and the Chinese Language*, pages 11–21.

Robert J. Glushko. 1979. The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 5(4):674. https://doi.org/10.1037/0096-1523.5.4.674

Janet Hui-wen Hsiao and Richard Shillcock. 2004. Connectionist modeling of Chinese character pronunciation based on foveal splitting. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26.

Janet Hui-wen Hsiao and Richard Shillcock. 2005. Differences of split and non-split architectures emerged from modelling chinese character pronunciation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 27.

Janet Hui-wen Hsiao and Richard Shillcock. 2006. Analysis of a chinese phonetic compound database: Implications for orthographic processing. *Journal of Psycholinguistic Research*, 35(5):405–426. https://doi.org/10.1007/s10936-006-9022-y, PubMed: 16897357

Chun-Hsien Hsu, Jie-Li Tsai, Chia-Ying Lee, and Ovid J.-L. Tzeng. 2009. Orthographic combinability and phonological consistency effects in reading Chinese phonograms: an event-related potential study. *Brain and Language*, 108(1):56–66. https://doi.org/10.1016/j.bandl.2008.09.002, PubMed: 18951624

Chih-Wei Hue. 1992. Recognition processes in character naming. In *Advances in Psychology*, volume 90, pages 93–107. Elsevier.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*.

Ho Cheong Lam. 2008. An exploratory study of the various ways that children read and write unknown Chinese characters. *Journal of Basic Education*, 17(1).

Ho Cheong Lam. 2014. Elaborating the concepts of part and whole in variation theory: The case of learning chinese characters. *Scandinavian Journal of Educational Research*, 58(3):337–360. https://doi.org/10.1080/00313831.2012.732604

Min Li, Marina Danilevsky, Sara Noeman, and Yunyao Li. 2018. Dimsim: An accurate chinese phonetic similarity algorithm based on learned high dimensional encoding. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 444–453. https://doi.org/10.18653/v1/K18-1043

Y. Li and J. S. Kang. 1993. Analysis of phonetics of the ideophonetic characters in modern Chinese. *Information Analysis of Usage of Characters in Modern Chinese*, pages 84–98.

Yunn-Wen Lien. 1985. Consistency of the phonetic clues in the Chinese phonograms and their naming latencies. *Psychological Department. National Taiwan University, Taipei.*

In-Mao Liu, S. C. Chen, and I. R. Sue. 2003. Regularity and consistency effects in chinese character naming. *Chinese Journal of Psychology*, 45(1):29–46.

Boning Lyu, Chun Lai, Chin-Hsi Lin, and Yang Gong. 2021. Comparison studies of typing and handwriting in chinese language learning: A synthetic review. *International Journal of Educational Research*, 106:101740. https://doi.org/10.1016/j.ijer.2021.101740

Gary F. Marcus, Ursula Brinkmann, Harald Clahsen, Richard Wiese, and Steven Pinker. 1995. German inflection: The exception that proves the rule. *Cognitive Psychology*, 29(3):189–256. https://doi.org/10.1006/cogp.1995.1015, PubMed: 8556846

Kate McCurdy, Sharon Goldwater, and Adam Lopez. 2020. Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for german plurals. *arXiv preprint arXiv:2005.08826*. https://doi.org/10.18653/v1/2020.acl-main.159

Tomohiko Morioka. 2008. Chise: Character processing based on character ontology. In *International Conference on Large-Scale Knowledge Resources*, pages 148–162. Springer. https://doi.org/10.1007/978-3-540-78159-2_14

David C. Plaut. 2005. Connectionist approaches to reading. *The Science of Reading: A handbook*, pages 24–38. https://doi.org/10.1002/9780470757642.ch2

David C. Plaut, James L. McClelland, Mark S. Seidenberg, and Karalyn Patterson. 1996. Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103(1):56. https://doi.org/10.1037/0033-295X.103.1.56, PubMed: 8650300

Mark S. Seidenberg and James L. McClelland. 1989. A distributed, developmental model of word recognition and naming. *Psychological Review*, 96(4):523. https://doi.org/10.1037/0033-295X.96.4.523, PubMed: 2798649

Li Hai Tan, Min Xu, Chun Qi Chang, and Wai Ting Siok. 2013. China's language input system in the digital age affects children's reading development. *Proceedings of the National Academy of Sciences*, 110(3):1119–1123. `https://doi.org/10.1073/pnas.1213586110`, PubMed: 23277555

Jie-Li Tsai, Erica Chung-I Su, Ovid J. L. Tzeng, and Daisy L. Hung. 2005. Consistency, regularity, and frequency effects in naming Chinese characters. *Language and Linguistics*, 6:75–107.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Endong Xun, Gaoqi Rou, Xiaoyue Xiao, and Jiaojiao Zhang. 2016. 大数据背景下 bcc 语料库的研制 [the construction of the bcc corpus in the age of big data]. 语料库语言学 *[Corpus Linguistics]*, 3(1):93–118.

Jianfeng Yang, Bruce D. McCandliss, Hua Shu, and Jason D. Zevin. 2009. Simulating language-specific and language-general effects in a statistical learning model of Chinese reading. *Journal of Memory and Language*, 61(2):238–257. `https://doi.org/10.1016/j.jml.2009.05.001`, PubMed: 20161189

Su-Ling Yeh, Wei-Lun Chou, and Pokuan Ho. 2017. Lexical processing of chinese sub-character components: Semantic activation of phonetic radicals as revealed by the stroop effect. *Scientific Reports*, 7(1):1–12. `https://doi.org/10.1038/s41598-017-15536-w`, PubMed: 29150618

Han Yuan, Eliane Segers, and Ludo Verhoeven. 2022. The role of phonological awareness, pinyin letter knowledge, and visual perception skills in kindergarteners' Chinese character reading. *Behavioral Sciences*, 12(8):254. `https://doi.org/10.3390/bs12080254`, PubMed: 36004825

Enlai Zhou. 1958. 当前文字改革的任务 [current tasks for writing system reform]. *Retrieved December 2, 2022 from https://www.marxists.org/chinese/zhouenlai/129.htm.*

Xiaolin Zhou and William Marslen-Wilson. 1999. Sublexical processing in reading chinese. In *Reading Chinese Script*, pages 49–76. Psychology Press.

Johannes C. Ziegler, Li Hai Tan, Conrad Perry, and Marie Montant. 2000. Phonology matters: The phonological frequency effect in written chinese. *Psychological Science*, 11(3):234–238. `https://doi.org/10.1111/1467-9280.00247`, PubMed: 11273409

## Appendix A. Chinese Character Naming Experiment

The human Chinese character naming experiment received IRB approval. The participants were recruited online and in person. The selection criteria include: 1) native speaker of Mandarin; 2) able to read and write in traditional Chinese scripts and pinyin. The participants completed an online questionnaire on Qualtrics on their phones or computers.

The participants were first asked to complete the screening questions to make sure that they are able to read and write in traditional Chinese scripts and pinyin. The screen questions are:

- 請將下面一段漢語拼音翻譯成漢字（聲調用數字代替）(Please transcribe the following pinyin into Chinese. Use numbers to represent the tone):
  ‘chun1 mian2 bu4 jue2 xiao3, chu4 chu4 wen2 ti2 niao3.’
- 請將下面一段漢字翻譯成漢語拼音（聲調用數字代替）(Please transcribe the following Chinese into pinyin. Use numbers to represent the tone):
  ‘夜來風雨聲, 花落知多少。’

Then the participants were asked to provide the pinyin for 60 test characters. The participants first selected ‘yes’ or ‘no’ whether they know the character. Then they were asked to type the pinyin of the character. Example questions are:

- 請問您認識" 紑" 這個字嗎? (Do you know the character 紑?)
  □ 認識 (yes) □ 不認識 (no)
- 請猜測並標註" 紑" 的讀音：(Please guess and write the pinyin of 紑)

The 60 test characters were separated into 2 test blocks, with 30 characters each. In between the 2 blocks, we set a block of 15 frequent characters and ask the participants to provide the answer to make it more engaging for the participants. An example question is:

- 請標註“河”的讀音: (Please write the pinyin of ‘‘河’’)

## Appendix B. Tables of Statistic Summaries

| data | label | –T–S | –T+S | +T–S | +T+S |
|---|---|---|---|---|---|
| ALL | BASE | 49.5 | 50.3 | 51.0 | 49.2 |
| | LABEL$_m$ | 48.2 | 49.0 | 49.8 | 48.8 |
| | LABEL$_s$ | 48.2 | 50.8 | 50.8 | 51.2 |
| | LABEL$_{mr}$ | 52.3 | 53.5 | 53.7 | 50.0 |
| | LABEL$_{sr}$ | 51.0 | 51.0 | 52.3 | 49.2 |
| MID | BASE | 47.0 | 43.7 | 45.3 | 46.7 |
| | LABEL$_m$ | 47.7 | 48.0 | 48.0 | 45.0 |
| | LABEL$_s$ | 49.7 | 49.3 | 48.7 | 43.0 |
| | LABEL$_{mr}$ | 46.0 | 45.0 | 45.3 | 47.7 |
| | LABEL$_{sr}$ | 49.0 | 52.0 | 49.7 | 45.3 |
| HIGH | BASE | 40.0 | 41.3 | 39.7 | 41.3 |
| | LABEL$_m$ | 39.3 | 39.7 | 42.0 | 39.0 |
| | LABEL$_s$ | 40.7 | 40.0 | 41.7 | 42.7 |
| | LABEL$_{mr}$ | 45.0 | 43.0 | 43.0 | 41.7 |
| | LABEL$_{sr}$ | 44.0 | 43.7 | 42.0 | 44.0 |
| ALL+ FREQ | BASE | 48.0 | 49.7 | 52.3 | 49.3 |
| | LABEL$_m$ | 46.0 | 47.3 | 48.0 | 49.3 |
| | LABEL$_s$ | 47.3 | 52.7 | 53.3 | 51.0 |
| | LABEL$_{mr}$ | 52.3 | 55.0 | 53.7 | 50.3 |
| | LABEL$_{sr}$ | 49.7 | 49.3 | 50.7 | 47.3 |

Table 11: The average accuracy (over 5 seeds) on the test set for models in Experiment 2 (MODEL[+PINYIN]) trained on HIGH, MID, or adding frequency label as input features on ALL. +T, –T, +S, –S refers to adding tone, no tone, shuffling, and no shuffling, respectively.

| Label | Freq. | # of characters | *regular* (%) | *alliterating* (%) | *rhyming* (%) | *irregular* (%) |
|---|---|---|---|---|---|---|
| Rare | = 1 | 1025 | 41.6 | 7.0 | 22.7 | 28.7 |
| Low | 2 – 29 | 1116 | 42.7 | 8.0 | 24.8 | 24.5 |
| Mid | 30 – 2337 | 1070 | 44.6 | 7.5 | 24.1 | 23.8 |
| High | $> 2337$ | 1070 | 42.1 | 8.7 | 22.5 | 26.7 |

Table 12: The summary of characters in ALL+FREQ.

| Overlap Rate | | No Tone | | | | Tone | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | No Shuffle | | Shuffle | | No Shuffle | | Shuffle | |
| Data | Model | MODEL [–PINYIN] | MODEL [+PINYIN] | MODEL [–PINYIN] | MODEL [+PINYIN] | MODEL [–PINYIN] | MODEL [+PINYIN] | MODEL [–PINYIN] | MODEL [+PINYIN] |
| ALL | BASE | 0.43±0.06 | 0.47±0.06 | 0.44±0.06 | 0.45±0.06 | 0.41±0.06 | 0.47±0.06 | 0.42±0.06 | 0.48±0.06 |
| | LABEL$_m$ | 0.41±0.06 | 0.49±0.06 | 0.43±0.05 | 0.49±0.07 | 0.42±0.05 | 0.49±0.07 | 0.42±0.06 | 0.48±0.06 |
| | LABEL$_{mr}$ | 0.43±0.07 | 0.45±0.05 | 0.43±0.05 | 0.47±0.06 | 0.46±0.06 | 0.43±0.06 | 0.44±0.06 | 0.45±0.07 |
| | LABEL$_s$ | 0.42±0.06 | 0.47±0.06 | 0.44±0.06 | 0.45±0.06 | 0.44±0.06 | 0.48±0.06 | 0.45±0.06 | 0.48±0.06 |
| | LABEL$_{sr}$ | 0.42±0.07 | 0.46±0.07 | 0.42±0.06 | 0.47±0.06 | 0.44±0.06 | 0.48±0.06 | 0.44±0.05 | 0.47±0.06 |
| MID | BASE | 0.41±0.06 | 0.46±0.07 | 0.41±0.06 | 0.44±0.06 | 0.39±0.06 | 0.45±0.07 | 0.41±0.06 | 0.45±0.07 |
| | LABEL$_m$ | 0.42±0.06 | 0.46±0.06 | 0.39±0.06 | 0.47±0.06 | 0.42±0.07 | 0.47±0.06 | 0.41±0.06 | 0.44±0.07 |
| | LABEL$_{mr}$ | 0.42±0.06 | 0.44±0.08 | 0.41±0.06 | 0.47±0.06 | 0.41±0.06 | 0.45±0.06 | 0.42±0.06 | 0.46±0.06 |
| | LABEL$_s$ | 0.41±0.05 | 0.48±0.06 | 0.42±0.06 | 0.47±0.07 | 0.37±0.06 | 0.48±0.07 | 0.40±0.07 | 0.46±0.07 |
| | LABEL$_{sr}$ | 0.38±0.07 | 0.45±0.08 | 0.42±0.06 | 0.47±0.06 | 0.41±0.06 | 0.45±0.06 | 0.39±0.06 | 0.45±0.06 |
| HIGH | BASE | 0.32±0.06 | 0.38±0.07 | 0.31±0.06 | 0.39±0.06 | 0.30±0.05 | 0.41±0.08 | 0.32±0.06 | 0.42±0.06 |
| | LABEL$_m$ | 0.32±0.06 | 0.41±0.08 | 0.30±0.06 | 0.40±0.06 | 0.31±0.06 | 0.42±0.06 | 0.32±0.06 | 0.39±0.07 |
| | LABEL$_{mr}$ | 0.34±0.07 | 0.45±0.07 | 0.32±0.06 | 0.43±0.08 | 0.33±0.07 | 0.45±0.07 | 0.32±0.06 | 0.44±0.07 |
| | LABEL$_s$ | 0.30±0.07 | 0.41±0.07 | 0.33±0.06 | 0.41±0.07 | 0.31±0.06 | 0.43±0.06 | 0.34±0.06 | 0.43±0.07 |
| | LABEL$_{sr}$ | 0.31±0.05 | 0.45±0.06 | 0.32±0.06 | 0.43±0.06 | 0.31±0.05 | 0.44±0.06 | 0.32±0.06 | 0.44±0.07 |
| ALL+FREQ | BASE | 0.43±0.06 | 0.46±0.07 | 0.42±0.06 | 0.45±0.07 | 0.44±0.06 | 0.47±0.06 | 0.43±0.06 | 0.47±0.06 |
| | LABEL$_m$ | 0.44±0.06 | 0.44±0.07 | 0.44±0.06 | 0.46±0.06 | 0.43±0.06 | 0.47±0.06 | 0.40±0.06 | 0.47±0.07 |
| | LABEL$_{mr}$ | 0.44±0.06 | 0.45±0.07 | 0.42±0.05 | 0.44±0.06 | 0.43±0.06 | 0.43±0.05 | 0.43±0.06 | 0.47±0.06 |
| | LABEL$_s$ | 0.43±0.06 | 0.46±0.06 | 0.44±0.06 | 0.48±0.06 | 0.42±0.06 | 0.50±0.06 | 0.41±0.06 | 0.48±0.06 |
| | LABEL$_{sr}$ | 0.43±0.05 | 0.46±0.06 | 0.43±0.06 | 0.47±0.06 | 0.41±0.07 | 0.46±0.06 | 0.45±0.06 | 0.45±0.06 |

Table 13: The overlap rate averaged over 275 pairs of answers (5 random seeds x 55 participants) for each model with different labels and conditions.

| | | Phonetic Radical | | | | Accuracy | |
|---|---|---|---|---|---|---|---|
| char. | correct pinyin | ortho. | pinyin | saliency | human<br>N=55 | MODEL<br>[-PINYIN]<br>N=400 | MODEL<br>[+PINYIN]<br>N=400 |
| 紑 | fou | 不 | bu | 0 | 0% | 1% | 0% |
| 汨 | rou, niu, nv | 丑 | chou | 0 | 16% | 81% | 63% |
| 飵 | zuo, ze, zha | 乍 | zha | 0.29 | 82% | 86% | 86% |
| 氿 | gui, jiu | 九 | jiu | 0.17 | 47% | 9% | 33% |
| 阤 | tuo | 也 | ye | 0 | 7% | 0% | 0% |
| 伃 | yu | 予 | yu | 0.33 | 84% | 40% | 41% |
| 秔 | geng, jing | 亢 | kang | 0.43 | 0% | 0% | 0% |
| 恔 | xiao, jiao | 交 | jiao | 0.63 | 96% | 96% | 97% |
| 汵 | gan, cen, han | 今 | jin | 0.31 | 18% | 0% | 0% |
| 軵 | rong, fu | 付 | fu | 1.00 | 98% | 100% | 100% |
| 軨 | ling | 令 | ling | 0.82 | 84% | 98% | 99% |
| 艅 | yu | 余 | yu | 0.14 | 64% | 8% | 12% |
| 翫 | wan | 元 | yuan | 0.33 | 20% | 24% | 21% |
| 姺 | shen, xian | 先 | xian | 0.20 | 69% | 43% | 85% |
| 侊 | guang | 光 | guang | 0.38 | 62% | 39% | 73% |
| 汭 | rui | 內 | nei | 0 | 5% | 2% | 2% |
| 溓 | lian, nian, xian | 兼 | jian | 0.12 | 24% | 47% | 49% |
| 鉋 | bao, pao | 包 | bao | 0.25 | 96% | 99% | 100% |
| 粘 | tian, shan, qian | 占 | zhan | 0.13 | 0% | 9% | 19% |
| 艩 | gui | 危 | wei | 0.36 | 22% | 51% | 40% |
| 訽 | gou | 句 | ju | 0.26 | 22% | 13% | 11% |
| 抲 | he, qia | 可 | ke | 0.31 | 7% | 46% | 43% |
| 咍 | hai | 台 | tai | 0.21 | 0% | 0% | 0% |
| 觡 | ge | 各 | ge | 0.25 | 44% | 28% | 33% |
| 佮 | ge | 合 | he | 0.24 | 2% | 2% | 3% |
| 蛣 | jie | 吉 | ji | 0.15 | 36% | 67% | 62% |
| 梠 | lv | 呂 | lv | 0.67 | 91% | 48% | 38% |
| 硞 | que, ke, ku | 告 | gao | 0.27 | 5% | 3% | 1% |
| 鎕 | tang | 唐 | tang | 1.00 | 96% | 98% | 97% |
| 飪 | ren | 壬 | ren | 1.00 | 84% | 91% | 99% |
| 煥 | huan | 奐 | huan | 1.00 | 96% | 69% | 90% |
| 羜 | zhu | 宁 | ning | 0.17 | 0% | 61% | 41% |
| 拕 | tuo | 它 | ta | 0 | 82% | 88% | 79% |
| 愃 | xuan | 宣 | xuan | 1.00 | 96% | 98% | 100% |
| 粔 | ju | 巨 | ju | 0.88 | 84% | 97% | 99% |
| 奸 | gan | 干 | gan | 0.43 | 29% | 32% | 56% |
| 靉 | ai | 愛 | ai | 1.00 | 87% | 24% | 72% |
| 猵 | bian, pian | 扁 | bian | 0.57 | 95% | 99% | 99% |
| 諟 | shi, di | 是 | shi | 0.08 | 22% | 12% | 17% |
| 咮 | zhou | 朱 | zhu | 0.83 | 0% | 0% | 0% |
| 鍛 | duan | 段 | duan | 1.00 | 67% | 37% | 80% |
| 鋂 | mei, meng | 每 | mei | 0.20 | 38% | 16% | 62% |
| 紬 | chou | 由 | you | 0.46 | 2% | 8% | 8% |
| 膰 | fan, pan | 番 | fan | 0.56 | 96% | 84% | 95% |
| 憕 | cheng,deng, zheng | 登 | deng | 0.64 | 96% | 99% | 99% |
| 袙 | pa | 白 | bai | 0.10 | 5% | 0% | 0% |
| 柀 | bi | 皮 | pi | 0.18 | 2% | 16% | 16% |
| 搤 | e | 益 | yi | 0.56 | 2% | 2% | 1% |
| 謓 | chen | 真 | zhen | 0.18 | 24% | 0% | 7% |
| 啌 | xiang, qiang | 空 | kong | 0.60 | 24% | 13% | 6% |
| 穋 | lu | 翏 | liu | 0.19 | 0% | 23% | 14% |
| 誚 | qiao | 肖 | xiao | 0.31 | 33% | 23% | 14% |
| 慔 | mo, mu | 莫 | mo | 0.82 | 98% | 98% | 97% |
| 鑗 | meng | 蒙 | meng | 1.00 | 96% | 91% | 100% |
| 黵 | dan | 詹 | zhan | 0.17 | 29% | 67% | 63% |
| 綌 | xi | 谷 | gu | 0 | 2% | 1% | 0% |
| 嗹 | lian | 連 | lian | 1.00 | 89% | 72% | 91% |
| 琲 | bei | 非 | fei | 0.67 | 0% | 0% | 1% |
| 禡 | ma | 馬 | ma | 0.75 | 80% | 28% | 76% |
| 螝 | hui, gui | 鬼 | gui | 0.18 | 62% | 42% | 60% |

Table 14: The humans' and models' results for each test character.